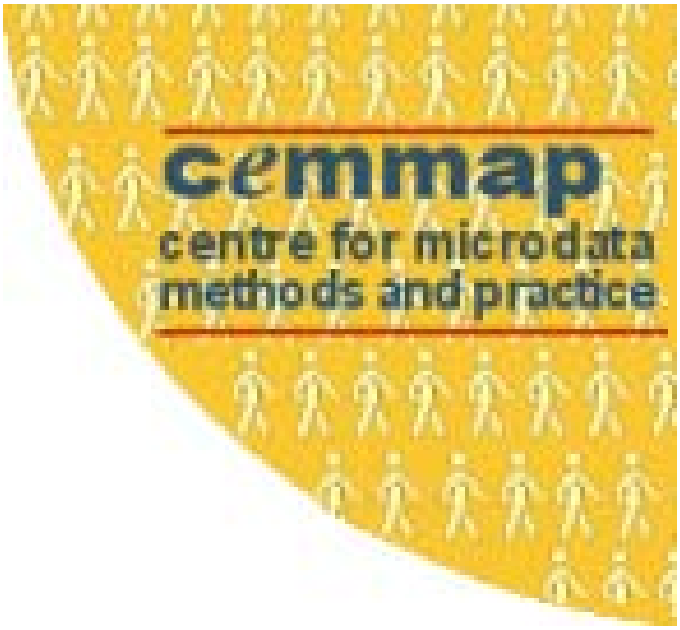

CHARACTERIZATION OF THE ASYMPTOTIC DISTRIBUTION OF SEMIPARAMETRIC M-ESTIMATORS

Hidehiko Ichimura
Sokbae Lee

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP15/06

Characterization of the Asymptotic Distribution of Semiparametric M-Estimators*

Hidehiko Ichimura

Graduate School of Public Policy and Graduate School of Economics
University of Tokyo

and

Sokbae Lee

Department of Economics
University College London

Email: ichimura@e.u-tokyo.ac.jp; l.simon@ucl.ac.uk

21 May 2006

Abstract

This paper develops a concrete formula for the asymptotic distribution of two-step, possibly non-smooth semiparametric M-estimators under general misspecification. Our regularity conditions are relatively straightforward to verify and also weaker than those available in the literature. The first-stage nonparametric estimation may depend on finite dimensional parameters. We characterize: (1) conditions under which the first-stage estimation of nonparametric components do not affect the asymptotic distribution, (2) conditions under which the asymptotic distribution is affected by the derivatives of the first-stage nonparametric estimator with respect to the finite-dimensional parameters, and (3) conditions under which one can allow non-smooth objective functions. Our framework is illustrated by applying it to three examples: (1) profiled estimation of a single index quantile regression model, (2) semiparametric least squares estimation under model misspecification, and (3) a smoothed matching estimator.

*We would like to thank Songnian Chen, Xiaohong Chen, Jinyong Hahn, and participants at numerous seminars for helpful comments. Also, we would like to thank the Leverhulme Trust and the Economic and Social Research Council (ESRC) through the funding of the Centre for Microdata Methods and Practice (*http://cemmap.ifs.org.uk*), of the research programme *Evidence, Inference and Inquiry* (*http://www.evidencescience.org*), and of ESRC Research Grant RES-000-22-0704, and the Japanese Government for Grants-in-Aid for Scientific Research. Part of this research is carried out while Lee was visiting the University of Tokyo. We thank the Graduate School of Economics at the University of Tokyo for its financial support through the COE Project and hospitality.

1 Introduction

This paper develops a concrete formula for the asymptotic distribution of two-step, possibly non-smooth semiparametric M-estimators under general misspecification. In particular, we obtain a direct way of characterizing the asymptotic distribution of two-step semiparametric M-estimators for which the first-stage nonparametric estimators may depend on unknown finite-dimensional parameters. In addition, we allow for smooth and non-smooth objective functions.

Our paper is closely related with Andrews (1994a), Newey (1994), Pakes and Olley (1995), Chen and Shen (1998), Ai and Chen (2003), Chen, Linton, and Van Keilegom (2003), and Chen (2005). Previous papers develop general forms to compute the asymptotic distribution of semiparametric estimators.

In this paper, we go a step further and characterize the asymptotic variance formula. We first characterize conditions under which the first-stage estimation of nonparametric components do not affect the asymptotic distribution. Andrews (1994a) and Newey (1994) have first derived sufficient conditions in the context of a smooth semiparametric GMM framework. Our results are applicable to semiparametric M-estimators with possibly non-smooth objective functions. Our results also provide a unifying interpretation of two apparently different results of Newey (1994, Propositions 2 and 3).

All the existing semiparametric estimators we have examined have asymptotic distributions unaffected by the derivatives of the first-stage nonparametric estimators with respect to the finite-dimensional parameters. We show that this is not the most general case and characterize conditions under which the derivatives do affect the asymptotic distribution.

We also characterize conditions under which one can allow non-smooth objective functions. When the nonparametric component depends on the finite-dimensional parameters, we require that the objective function have a linear representation with respect to both parametric and nonparametric components with regularity conditions on the remainder term. When the nonparametric component does not depend on the finite-dimensional parameters, the objective function can be less smooth with respect to the nonparametric part. We also require that the first-stage nonparametric estimator be differentiable with respect to finite-dimensional parameters asymptotically.

Our approach is analogous to the standard analysis of the two-step parametric estimators when the objective function is not smooth. To be more specific, our approach is based on a Taylor series expansion of the expectation of the objective function (see, e.g. Pollard (1984) and Sherman (1994)). Since the first stage involves nonparametric estimation and thus the

objective function is a functional defined on the Cartesian product of a Euclidean space and a function space, we need to use basic results of functional analysis and also need to modify the concept of asymptotic linearity of the first-stage nonparametric estimator suitably.[1] As a result, calculating a formula for the asymptotic distribution involves Fréchet differentiation of the expectation of an objective function. For many leading examples, this is often easy to derive (see, e.g. Ichimura (2006)).

To establish the asymptotic theory, we make the use of the idea behind Pollard (1984, pp.140-142) and apply empirical process methods of Van der Vaart and Wellner (1996) to deal with remainder terms in the asymptotic expansion of the objective function. A more common practice of using stochastic equicontinuity used by e.g, Andrews (1994a,b), Newey (1994), Chen, Linton, and Van Keilegom (2003), and Ichimura (2006) is applicable to semiparametric GMM estimators but is not directly applicable to semiparametric M-estimators.

Our framework is illustrated by applying it to profiled estimation of a single index quantile regression model. Due to the nature of profiled estimation and non-differentiability of the check function it is non-trivial to analyze this estimator. Our general framework allows us to calculate the asymptotic distribution of this estimator. Our framework is also illustrated by applying it to semiparametric least squares estimation of Ichimura (1993) under model misspecification. We show that while the first stage estimation does not affect the asymptotic distribution regardless of whether the model is misspecified or not, the asymptotic distribution is different under the two cases. The result of the latter example can be viewed as a semiparametric analog of White (1981), who characterizes the asymptotic distribution of parametric least squares for misspecified nonlinear regression models. To the best of our knowledge, both of these results are new findings in the literature. Finally, the paper considers a smoothed matching estimator to illustrate the effects of first-stage estimation. This example shows the simple nature of the form of the correction term in our characterization.

The paper is organized as follows. Section 2 defines a semiparametric M-estimator and describes examples. Section 3 provides theoretical results, including regularity conditions and general formulas for the asymptotic distribution. Section 4 demonstrates usefulness of the main results of Section 3 by applying them to all the aforementioned examples. All the proofs are in the Appendix.

[1] As an early paper that uses the functional analysis approach in econometrics, Aït-Sahalia (1994) develops a generalized delta method for functionals of nonparametric kernel estimators using functional derivatives.

2 Estimation

Suppose that there exists a vector of finite-dimensional parameters θ_0 that minimizes $E[m(Z, \theta, f_0(\cdot, \theta))]$ for an unknown, d_f-vector-valued function f_0, where $m(Z, \theta, f_0(\cdot, \theta))$ is a known, real-valued function of data $Z \in \mathbf{R}^{d_z}$ and θ directly and indirectly through f_0. Assume that $f_0(\cdot, \theta)$ is a function of Z, possibly indexed by θ. For simplicity in notation, the arguments of f_0 are denoted here by a dot. This notation is useful because we can allow $m(Z, \theta, f_0(\cdot, \theta))$ to depend either on the whole function $f_0(\cdot, \theta)$ or on values of $f_0(\cdot, \theta)$ at some data points.

Throughout the paper, let $\theta \in \Theta$ denote finite dimensional parameters, where Θ is a compact subset of $\mathbf{R}^{d_\theta}$, and for each θ, let $f(\cdot, \theta) \in \mathcal{F}$ denote infinite dimensional parameters, where $\mathcal{F}$ is a Banach space with the supremum norm.[2] More concretely, the parameter space $\Theta \times \mathcal{F}$ is a Cartesian product of Θ and $\mathcal{F}$ with a norm defined by $\|(\theta, f)\|_{\Theta \times \mathcal{F}} = \|\theta\| + \|f\|_{\mathcal{F}}$, where $\|\theta\|$ is the usual matrix norm and $\|f\|_{\mathcal{F}} = \sup_{\theta \in \Theta} \sup_{z \in \mathcal{S}} \|f(z, \theta)\|$ for any $f \in \mathcal{F}$, where $\mathcal{S}$ is a subset of the support of the data Z.[3] We will use the notation $\|\cdot\|_\infty$ to denote the supremum norm. When f depends on θ, $\|f(\cdot, \theta)\|_\infty$ will be understood as the supremum norm with θ fixed. Thus, $\|f\|_{\mathcal{F}} = \sup_{\theta \in \Theta} \|f(\cdot, \theta)\|_\infty$.

Assume that for each θ, a nonparametric estimator $\hat{f}_n(\cdot, \theta)$ of $f_0(\cdot, \theta)$ is available. Furthermore, assume that the observed data $\{Z_i : i = 1, \ldots, n\}$ are a random sample of Z. A natural sample analog estimator of θ_0 is an M-estimator that minimizes

$$\hat{S}_n(\theta) \equiv n^{-1} \sum_{i=1}^{n} m(Z_i, \theta, \hat{f}_n(\cdot, \theta)). \tag{2.1}$$

Let $\hat{\theta}_n$ denote the resulting estimator of θ_0.

There are many examples of semiparametric estimators that can be viewed as special cases of (2.1). Some well-known examples include: Robinson (1988), Powell, Stock, and Stoker (1989), Ichimura (1993), and Klein and Spady (1993) among many others. To illustrate the main result of this paper, we will analyze the following three examples.

Example 2.1. *Profiled Estimation of a Single-Index Quantile Regression Model.* This model has the form

$$Y = G_0(X_1 + X_2^T \theta_0) + U, \tag{2.2}$$

[2] Instead of the supremum norm, one may develop results parallel to those obtained in this paper using a different norm, say the L_2 norm.

[3] In examples considered in the paper, $\mathcal{S}$ is the intersection of the support of the data and the support of the trimming function. This is due to the usual technical reason regarding the first-stage kernel estimation.

where Y is the dependent variable, $X = (X_1, X_2) \in \mathbf{R}^{d_x}$ is a vector of explanatory variables, θ_0 is a vector of unknown parameters, $G_0(\cdot)$ is an unknown, real-valued function, and the τ-quantile of U given $X = x$ is zero for almost every x for some τ, $0 < \tau < 1$. Here, T denotes a transpose. To guarantee identification, we assume that X_1 is continuously distributed and its coefficient is non-zero and is normalized to be one.

To describe our estimator of θ_0, let $\rho_\tau(u)$ denote the 'check' function, that is $\rho_\tau(u) = |u| + (2\tau - 1)u$, let $f_0(t, \theta)$ denote the τ-quantile of Y conditional on $X_1 + X_2'\theta = t$ for each θ and on the event that $X \in \mathcal{T}$ with a known compact set $\mathcal{T}$, and let $\hat{f}_n(t, \theta)$ denote a nonparametric estimator of $f_0(t, \theta)$. Then $G_0(x_1 + x_2\theta_0) = f_0(x_1 + x_2\theta_0, \theta_0)$. In principle, any reasonable nonparametric estimator could be used, as long as a nonparametric estimator satisfies some regularity conditions, which will be given in Section 3. To be specific, $\hat{f}_n(X_{1i} + X_{2i}^T\theta, \theta)$ is defined as a smoothed local linear quantile regression estimator (Chaudhuri (1991)), that is $\hat{f}_n(X_{1i} + X_{2i}^T\theta, \theta) \equiv \hat{c}_{ni}(\theta)$, where $\hat{c}_{ni}(\theta) \equiv [\hat{c}_{ni0}(\theta), \hat{c}_{ni1}(\theta)]'$ solves the following minimization problem

$$\min_{(c_0,c_1)\in\mathbf{R}^2} \sum_{j=1}^{n} 1(X_j \in \mathcal{T}_n)\tilde{\rho}_{\tau,n}\left[Y_j - c_0 - c_1(X_{1j} + X_{2j}^T\theta - X_{1i} - X_{2i}^T\theta)\right] \times K\left(\frac{X_{1j} + X_{2j}^T\theta - X_{1i} - X_{2i}^T\theta}{h_n}\right). \tag{2.3}$$

Here, $\tilde{\rho}_{\tau,n}$ is a smoothed version of $\rho_\tau(u)$ as in Horowitz (1998), $1(\cdot)$ is the usual indicator, $K(\cdot)$ is a kernel function, h_n is a sequence of bandwidths that converges to zero as $n \to \infty$, and $\mathcal{T}_n = \{x : B(x; 2h_n) \subset \mathcal{T}\}$, where $B(x, r)$ is a r-radius ball centered at x. The smoothed estimator is used here to ensure that $\hat{f}_n$ is Lipschitz continuous for both arguments with probability tending to one.

An estimator of θ_0 is now defined as

$$\hat{\theta}_n = \text{argmin}_\theta n^{-1} \sum_{i=1}^{n} 1(X_i \in \mathcal{T})\rho_\tau\left[Y_i - \hat{f}_n(X_{1i} + X_{2i}^T\theta, \theta)\right]. \tag{2.4}$$

As in Ichimura (1993), the trimming function $1(\cdot \in \mathcal{T})$ is necessary to ensure that the density of $X_1 + X_2^T\theta$ is bounded away from 0 on $\mathcal{T}$ for any θ.[4]

It is worth mentioning existing estimators of θ_0. Chaudhuri, Doksum, and Samarov (1997) developed average derivative estimators of θ_0 and Khan (2001) proposed a two-step rank estimator of θ_0. The new estimator is applicable to more general cases than

[4]One can use a more sophisticated trimming function that converges to one as $n \to \infty$. For example, Robinson (1988) uses the trimming function $1(\hat{p}(x) > c_n)$, where $\hat{p}(x)$ is the kernel density estimator of X and c_n is a sequence of positive real numbers converging to zero at a sufficiently slow rate. See Ichimura (2006).

the estimators of Chaudhuri, Doksum, and Samarov (1997) in the sense that X can include discrete variables and functionally dependent variables (e.g., the square of one of explanatory variables) and than the estimator of Khan (2001) in the sense that monotonicity of f_0 is not required.

To apply the general result obtained in the paper, let

$$m(z,\theta,f(\cdot,\theta)) = \frac{1}{2}1(x \in \mathcal{T})\rho_\tau\left[y - f(x_1 + x_2^T\theta,\theta)\right], \tag{2.5}$$

where $z = (y,x)$ and $x = (x_1,x_2)$. Our estimator $\hat{\theta}_n$ is an M-estimator in (2.1) with $m(z,\theta,f(\cdot,\theta))$ defined above.

Example 2.2. *Semiparametric Least Squares Estimation under Misspecification.* This example is concerned with the asymptotic distribution of the semiparametric least squares (SLS) estimator of Ichimura (1993) under model misspecification. Let $E_{\mathcal{T}}$ denote an expectation conditional on $X \in \mathcal{T}$. As in the previous example, we assume that for identification, there exists a continuously distributed component of $X = (X_1, X_2)$, say X_1, whose coefficient is non-zero and is normalized to be one. Let θ denote a vector of coefficients of X_2 and θ_0 denote the true value of θ in a sense that θ_0 minimizes

$$E\left[1(X \in \mathcal{T})\{Y - f_0(X_1 + X_2^T\theta,\theta)\}^2\right], \tag{2.6}$$

where $\mathcal{T}$ is a known compact set and $f_0(t,\theta)$ denotes the expectation of Y conditional on $X_1 + X_2^T\theta = t$ and on the event that $X \in \mathcal{T}$ for each θ. Therefore, under model misspecification, $f_0(x_1 + x_2^T\theta_0,\theta_0)$ can be interpreted as the best L^2 approximation to $E_{\mathcal{T}}[Y|X=x]$ in the class of single-index models since $f_0(X_1 + X_2^T\theta,\theta)$ is the best L^2 approximation to $E_{\mathcal{T}}[Y|X=x]$ for each fixed θ and (2.6) implies that θ_0 minimizes

$$E\left[1(X \in \mathcal{T})\{E_{\mathcal{T}}[Y|X=x] - f_0(X_1 + X_2^T\theta,\theta)\}^2\right]. \tag{2.7}$$

The SLS estimator of Ichimura (1993), say $\hat{\theta}_n$, minimizes a sample analog of (2.6). That is, $\hat{\theta}_n$ solves

$$\min_\theta n^{-1}\sum_{i=1}^n 1(X_i \in \mathcal{T})\left[Y_i - \hat{f}_n(X_{1i} + X_{2i}^T\theta,\theta)\right]^2, \tag{2.8}$$

where $\hat{f}_n(\cdot,\theta)$ is a nonparametric kernel estimator of $f_0(\cdot,\theta)$ defined in Ichimura (1993, p.78). The asymptotic distribution of the SLS estimator is established by Ichimura (1993) under the assumption that the model is correctly specified, that is $E_{\mathcal{T}}[Y|X=x] = f_0(x_1 +$

$x_2^T\theta_0, \theta_0)$. In this paper, we establish the asymptotic distribution of the SLS estimator when $E_{\mathcal{T}}[Y|X = x]$ may not belong to a class of single-index models.

Let

$$m(z, \theta, f(\cdot, \theta)) = \frac{1}{2} 1(x \in \mathcal{T}) \left[y - f(x_1 + x_2^T\theta, \theta)\right]^2, \tag{2.9}$$

where $z = (y, x)$ and $x = (x_1, x_2)$. The SLS estimator $\hat{\theta}_n$ is an M-estimator with $m(z, \theta, f(\cdot, \theta))$ defined above.

Example 2.3. *Smoothed Matching Estimator*. This example is concerned about estimating the average treatment on the treated, that is $\theta_0 = E[Y_1 - Y_0|D = 1, X \in \mathcal{T}]$, where Y_1 and Y_0 are potential outcomes and D is the treatment status (e.g., Heckman, Ichimura, and Todd (1998)). Note that the θ_0 is defined conditional on the event that $X \in \mathcal{T}$ for some compact set $\mathcal{T}$ over which both the densities of X given $D = 0$ and X given $D = 1$ are bounded away from 0. The main estimation problem in this example is to construct the counterfactual $E_{\mathcal{T}}[Y_0|D = 1]$. Suppose that $E_{\mathcal{T}}[Y_0|X, D = 1] = E_{\mathcal{T}}[Y_0|X, D = 0]$ for a high-dimensional X (ignorability assumption). Then one may use a kernel estimator of $f_0(x) = E_{\mathcal{T}}[Y_0|X = x, D = 0]$ with a trimming function as in Examples 2.1 and 2.2. Assume that $\{(Y_i, X_i, D_i) : i = 1, \ldots, n\}$ is a random sample of (Y, X, D), where $Y = DY_1 + (1 - D)Y_0$. Then an estimator $\hat{f}_n(x)$ can be defined as

$$\hat{f}_n(x) = [nh_n^{d_x} p_n(x)]^{-1} \sum_{i=1}^{n} 1(X_i \in \mathcal{T}_n)(1 - D_i)Y_{0i} K\left(\frac{x - X_i}{h_n}\right),$$

where $p_n(x) = (nh_n^{d_x})^{-1} \sum_{i=1}^{n} 1(X_i \in \mathcal{T}_n)(1 - D_i)K\left[(x - X_i)/h_n\right]$, K is a kernel function with a bandwidth h_n, and d_x is the dimension of X. A semiparametric estimator of θ_0 can be obtained by an M-estimator with

$$m(z, \theta, f(\cdot)) = \frac{1}{2} 1(x \in \mathcal{T}) d \left[\theta - (y_1 - f(x))\right]^2, \tag{2.10}$$

where $z = (y_1, d, x)$.[5]

3 Asymptotic Results

3.1 Assumptions

In this subsection, we state assumptions that are needed to establish asymptotic results. The consistency of a semiparametric M-estimator $\hat{\theta}_n$ can be obtained using general results

[5]Powell (1994) argues that this is a nonparametric formulation.

available in the literature. See, for example, Theorem 2.1 of Newey and McFadden (1994, p.2121), Corollary 3.2.3 of Van der Vaart and Wellner (1996, p.287), and Theorem 1 of Chen, Linton, and Van Keilegom (2003). Thus, we assume that $\hat{\theta}_n$ is consistent and consider only a neighborhood of θ_0. For any $\delta_1 > 0$ and $\delta_2 > 0$, define $\Theta_{\delta_1} = \{\theta \in \Theta : \|\theta - \theta_0\| < \delta_1\}$ and $\mathcal{F}_{\delta_1,\delta_2} = \{f \in \mathcal{F} : \sup_{\theta\in\Theta_{\delta_1}} \|f(\cdot,\theta) - f_0(\cdot,\theta_0)\|_\infty < \delta_2\}$. For any function ψ of data, let $\|\psi(Z)\|_{L^2(P)} = [\int[\psi(Z)]^2 dP]^{1/2}$, where P is the probability measure of data Z. That is, $\|\cdot\|_{L^2(P)}$ is the $L^2(P)$-norm. To simplify the notation, we assume in Section 3 that $d_f = 1$, i.e., $f(\cdot,\theta)$ is a real-valued function.[6]

To establish asymptotic results, we make the following assumptions:

Assumption 3.1. *(a) θ_0 is an interior point in Θ, which is a compact subset of $\mathbf{R}^{d_\theta}$.*

(b) θ_0 is a unique minimizer of $E[m(Z,\theta,f_0(\cdot,\theta))]$.

(c) $\hat{\theta}_n \to_p \theta_0$.

Condition (a) is standard, condition (b) imposes identification, and condition (c) assumes the consistency of $\hat{\theta}_n$ to θ_0 in probability.

Assumption 3.2. *For any (θ_1, f_1) and (θ_2, f_2) in $\Theta_{\delta_1} \times \mathcal{F}_{\delta_1,\delta_2}$, there exist linear operators $\Delta_1(z, \theta_1 - \theta_2)$ and $\Delta_2(z, f_1(\cdot) - f_2(\cdot))$ and a function $\dot{m}(z,\delta_1,\delta_2)$ satisfying*

$$(a)\ \ |m(z,\theta_1,f_1(\cdot)) - m(z,\theta_2,f_2(\cdot)) - \Delta_1(z,\theta_1-\theta_2) - \Delta_2(z,f_1(\cdot)-f_2(\cdot))| \\ \leq [\|\theta_1-\theta_2\| + \|f_1(\cdot)-f_2(\cdot)\|_\infty]\,\dot{m}(z,\delta_1,\delta_2),$$

and

$$(b)\ \ \|\dot{m}(Z,\delta_1,\delta_2)\|_{L^2(P)} \leq C\,(\delta_1^{\alpha_1} + \delta_2^{\alpha_2})$$

for some constants $C < \infty$, $\alpha_1 > 0$, and $\alpha_2 > 0$.[7]

Since Δ_1 is a linear operator and θ is a finite-dimensional parameter, we write $\Delta_1(z,\theta_1 - \theta_2) = \Delta_1(z)\cdot(\theta_1 - \theta_2)$. Assumption 3.2 allows for both differentiable and non-differentiable functions with respect to parameters.

[6] It is rather straightforward to extend our main result in Section 3 to a vector-valued $f(\cdot,\theta)$ with the use of more complicated notation. Appendix A presents the extension for the case $d_f > 1$.

[7] Here, Δ_1, Δ_2, and $\dot{m}$ may depend on $(\theta_2, f_2(\cdot))$. However, we suppress the dependence on $(\theta_2, f_2(\cdot))$ for the sake of simplicity in notation.

Example 2.1 Continued: To verify Assumption 3.2, define

$$\begin{aligned}\Delta_1(z) &= 0,\\ \Delta_2(z, f_1(\cdot) - f_2(\cdot)) &= -1(x \in \mathcal{T})[\tau - 1(y - f_2(x) \leq 0)](f_1(x) - f_2(x))\\ &\text{and}\\ \dot{m}(z, \delta_1, \delta_2) &= 1(x \in \mathcal{T})1(|y - f_2(x)| \leq \delta_2).\end{aligned}$$

As in Pollard (1991), note that

(3.1)
$$\begin{aligned}&\frac{1}{2}\Big|\rho_\tau[y - \{f(x,\theta) + h(x)\}] - \rho_\tau[y - f(x,\theta)] + 1(x \in \mathcal{T})[\tau - 1(y - f(x,\theta) \leq 0)][h(x)]\Big|\\ &\leq |h(x)|\, 1(x \in \mathcal{T})1\left\{|y - f(x,\theta)| \leq |h(x)|\right\}.\end{aligned}$$

Then since m depends on θ only through f, Assumption 3.2 (a) is satisfied by (3.1). To check Assumption 3.2 (b), assume that $|P_{Y|X}(y_1|x) - P_{Y|X}(y_2|x)| \leq C(x)|y_1 - y_2|$ for some function $C(x)$ such that $E[1(X \in \mathcal{T})C(X)] < \infty$, where $P_{Y|X}(\cdot|x)$ is the CDF of Y conditional on $X = x$. Then notice that

$$\begin{aligned}\|\dot{m}(Z, \delta_1, \delta_2)\|^2_{L^2(P)} &= E[1(X \in \mathcal{T})P_{Y|X}(f_2(\cdot) + \delta_2|X)] - E[1(X \in \mathcal{T})P_{Y|X}(f_2(\cdot) - \delta_2|X)]\\ &\leq C\delta_2\end{aligned}$$

for some positive constant C, implying that Assumption 3.2 (b) is satisfied with $\alpha_2 = 0.5$.

Examples 2.2 – 2.3 Continued: In these examples, Assumption 3.2 is trivially satisfied with $\alpha_1 = 1$ and $\alpha_2 = 1$.

Assumption 3.3. *Let $m^*(\theta, f) = E[m(Z, \theta, f)]$ for fixed θ and f. $m^*(\theta, f)$ is twice continuously Fréchet differentiable in an open, convex neighborhood of $(\theta_0, f_0(\cdot, \theta_0))$ with respect to a norm $\|(\theta, f)\|_{\Theta \times \mathcal{F}}$.*

This assumption implies that a second-order Taylor expansion of $m^*(\theta, f)$ is well defined.[8] Let $D_\theta m^*(\theta, f)$ and $D_f m^*(\theta, f)$ denote the partial Fréchet derivatives of $m^*(\theta, f)$ with respect to θ and f, respectively. In addition, let $D_{\theta\theta} m^*(\theta, f)$, $D_{\theta f} m^*(\theta, f)$, and

[8]Note that in Assumption 3.3, f is not indexed by θ. As a result, it is unnecessary to assume the differentiability of f with respect to θ in this expansion; however, it is needed to evaluate the Taylor expansion of $m^*(\theta, f)$ at $(\theta, f) = (\theta, f(\cdot, \theta))$. Hence, we assume that $f(\cdot, \theta)$ belongs to the common space $\mathcal{F}$ for any θ.

$D_{ff}m^*(\theta, f)$ denote second-order partial Fréchet derivatives of $m^*(\theta, f)$.[9] By Taylor's Theorem on Banach spaces (see, for example, Section 4.6 of Zeidler, 1986), if $m^*(\theta, f)$ is twice continuously Fréchet differentiable in an open, convex neighborhood of $(\theta_0, f_0(\cdot, \theta_0))$ with respect to a norm $\|(\theta, f)\|_{\Theta \times \mathcal{F}}$, then for any (θ, f) and (θ_0, f_0) in an open, convex neighborhood of $(\theta_0, f_0(\cdot, \theta_0))$,

$$
\begin{aligned}
(3.2) \qquad & m^*(\theta, f) - m^*(\theta_0, f_0) \\
& = D_\theta m^*(\theta_0, f_0)[\theta - \theta_0] + D_f m^*(\theta_0, f_0)[f - f_0] \\
& + \int_0^1 (1-s)\Big[D_{\theta\theta} m^*(\theta_s, f_s)[\theta - \theta_0, \theta - \theta_0] + 2D_{\theta f} m^*(\theta_s, f_s)[\theta - \theta_0, f - f_0] \\
& + D_{ff} m^*(\theta_s, f_s)[f - f_0, f - f_0]\Big]\, ds,
\end{aligned}
$$

where $\theta_s = \theta_0 + s(\theta - \theta_0)$ and $f_s = f_0 + s(f - f_0)$.

Example 2.1 Continued: Let $m^*(\theta, f) = E[m(Z, \theta, f)]$ for fixed θ and f, where $m(z, \theta, f)$ is defined in (2.5). First of all, since m depends on θ only through $f(\cdot, \theta)$,

$$
D_\theta m^*(\theta, f) = D_{\theta\theta} m^*(\theta, f) = D_{\theta f} m^*(\theta, f) \equiv 0.
$$

Use (3.1) to obtain

$$
\begin{aligned}
& |m^*(\theta, f + h) - m^*(\theta, f) + E[1(X \in \mathcal{T})\{\tau - 1(Y - f(X_1 + X_2^T\theta, \theta) \le 0)\}h(X)]| \\
& \le E[1\{|Y - f(X_1 + X_2^T\theta, \theta)| \le |h(X)|\}|h(X)|] \\
& \le E[1\{|Y - f(X_1 + X_2^T\theta, \theta)| \le |h(X)|\}]\, \|h\|_\infty \\
& = o(\|h\|_\infty)
\end{aligned}
$$

for any h in a neighborhood of zero. Thus,

$$
(3.3) \qquad D_f m^*(\theta, f)[h] = -E[1(X \in \mathcal{T})\{\tau - 1(Y - f(X_1 + X_2^T\theta, \theta) \le 0)\}h(X)].
$$

To compute $D_{ff}m^*(\theta, f)$, let $p_{Y|X}(y|x)$ denote the PDF of Y conditional on $X = x$. Notice that

$$
\begin{aligned}
& D_f m^*(\theta, f + h_2)[h_1] - D_f m^*(\theta, f)[h_1] \\
& = -E[1(X \in \mathcal{T})\{\tau - P_{Y|X}(f(X_1 + X_2^T\theta, \theta) + h_2(X)|X)\}h_1(X)] \\
& + E[1(X \in \mathcal{T})\{\tau - P_{Y|X}(f(X_1 + X_2^T\theta, \theta)|X)\}h_1(X)] \\
& = E[1(X \in \mathcal{T})p_{Y|X}(f(X_1 + X_2^T\theta, \theta)|X)h_2(X)h_1(X)] + o(\|h_2\|_\infty)
\end{aligned}
$$

[9] See monographs on nonlinear functional analysis such as Berger (1977) and Zeidler (1986) for well-established results of Fréchet differentiation in Banach spaces.

for any h_1 and h_2 in a neighborhood of zero. Thus,

$$D_{ff}m^*(\theta, f)[h_1, h_2] = E[1(X \in \mathcal{T})p_{Y|X}(f(X_1 + X_2^T\theta, \theta)|X)h_1(X)h_2(X)]. \tag{3.4}$$

Example 2.2 Continued: In this example, let $m^*(\theta, f) = E[m(Z, \theta, f)]$ for fixed θ and f, where $m(z, \theta, f)$ is defined in (2.9). As in Example 2.1, note that m depends on θ only through $f(\cdot, \theta)$. Hence,

$$\Delta_1(z) = D_\theta m^*(\theta, f) = D_{\theta\theta}m^*(\theta, f) = D_{\theta f}m^*(\theta, f) \equiv 0.$$

To compute $D_f m^*(\theta, f)[h]$, note that

$$m^*(\theta, f + h) - m^*(\theta, f) = -E[1(X \in \mathcal{T})\{Y - f(X_1 + X_2^T\theta, \theta))\}h(X)] + E[1(X \in \mathcal{T})h^2(X)]$$

for any h in a neighborhood of zero. Hence,

$$D_f m^*(\theta, f)[h] = -E[1(X \in \mathcal{T})\{Y - f(X_1 + X_2^T\theta, \theta))\}h(X)]. \tag{3.5}$$

To compute $D_{ff}m^*(\theta, f)[h_1, h_2]$, note that

$$D_f m^*(\theta, f + h_2)[h_1] - D_f m^*(\theta, f)[h_1] = E[1(X \in \mathcal{T})h_1(X)h_2(X)]$$

for any h_1 and h_2 in a neighborhood of zero. Therefore,

$$D_{ff}m^*(\theta, f)[h_1, h_2] = E[1(X \in \mathcal{T})h_1(X)h_2(X)].$$

Example 2.3 Continued: It is easy to show that

$$D_f m^*(\theta, f)[h] = E[1(X \in \mathcal{T})D\{\theta - (Y_1 - f(X))\}h(X)] \text{and} \tag{3.6}$$

$$D_{ff}m^*(\theta, f)[h_1, h_2] = E[1(X \in \mathcal{T})Dh_1(X)h_2(X)]. \tag{3.7}$$

To take account of the effect of the first-stage nonparametric estimation, it is necessary to consider a suitably-defined class of functions. In this paper, we consider a class of smooth functions defined in Van der Vaart and Wellner (1996, p.154), denoted by $\mathcal{C}_M^\alpha(\mathcal{X})$.[10] To be

[10] Although this class of functions seems to be quite general, in some applications, it may be more natural to use different classes of functions, e.g., a VC-class of functions, a class of monotone functions or that of convex functions. See Van der Vaart and Wellner (1996, in particular, Sections 2.6 and 2.7) for details for alternative classes of functions.

precise, we provide the exact definition of $\mathcal{C}_M^\alpha(\mathcal{X})$. Let $\underline{\alpha}$ denote the greatest integer strictly smaller than α, and for any vector $k = (k_1, \ldots, k_d)$ of d integers and let D^k denote the differential operator

$$D^k = \frac{\partial^{k.}}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}} \quad \text{with } k. = \sum_{i=1}^d k_i.$$

In addition, let

$$\|g\|_\alpha = \max_{k. \le \underline{\alpha}} \sup_x |D^k g(x)| + \max_{k. = \underline{\alpha}} \sup_{x,y} \frac{|D^k g(x) - D^k g(y)|}{\|x - y\|^{\alpha - \underline{\alpha}}},$$

where the suprema are taken over all x, y in the interior of $\mathcal{X}$ with $x \neq y$. Then $\mathcal{C}_M^\alpha(\mathcal{X})$ is the set of all continuous functions $g : \mathcal{X} \subset \mathbf{R}^d \mapsto \mathbf{R}$ with $\|g\|_\alpha \le M$.

Assumption 3.4. *(a) For any $\theta \in \Theta_{\delta_1}$, $f_0(\cdot, \theta)$ is an element of $\mathcal{C}_M^\alpha(\mathcal{X})$ for some $\alpha > d_1/2$, where d_1 is the dimension of the first argument of $f_0(\cdot, \theta)$ and $\mathcal{X}$ is a finite union of bounded, convex subsets of $\mathbf{R}^{d_1}$ with nonempty interior.*

(b) For any $\theta \in \Theta_{\delta_1}$, $\hat{f}_n(\cdot, \theta) \in \mathcal{C}_M^\alpha(\mathcal{X})$ with probability approaching one.

(c) $\sup_{\theta \in \Theta_{\delta_1}} \left\| \hat{f}_n(\cdot, \theta) - f_0(\cdot, \theta) \right\|_\infty = O_p(\tilde{\delta}_2)$ for $\tilde{\delta}_2$ satisfying $n^{1/2} \tilde{\delta}_2^{1+\alpha_2} \to 0$.

(d) As a function of θ, $f_0(\cdot, \theta)$ is twice continuously differentiable on Θ_{δ_1} with bounded derivatives on $\mathcal{X}$.

(e) For any $\varepsilon > 0$ and $\delta > 0$, independent of θ, there exists n_0 such that for all $n \ge n_0$, the following holds:

$$Pr\left\{ \left\| [\hat{f}_n(\cdot, \theta) - \hat{f}_n(\cdot, \theta_0)] - [f_0(\cdot, \theta) - f_0(\cdot, \theta_0)] \right\|_\infty \le \delta \|\theta - \theta_0\| \right\} \ge 1 - \varepsilon. \tag{3.8}$$

Condition (a) imposes smoothness condition on $f_0(\cdot, \theta)$ for each fixed θ. It is reasonable to assume that $f_0(\cdot, \theta)$ is a smooth function; however, a nonparametric estimator of $f_0(\cdot, \theta)$ may not share the same smoothness for fixed sample size n. Condition (b) assumes that a nonparametric estimator of $f_0(\cdot, \theta)$ shares the same smoothness condition with probability tending to one. Condition (c) requires some uniform rate of convergence of $\hat{f}_n(\cdot, \theta)$ in probability. If $\alpha_2 = 1$ (smooth m), $\tilde{\delta}_2 = o(n^{-1/4})$; when $\alpha_2 = 0.5$ (non-smooth m), $\tilde{\delta}_2 = o(n^{-1/3})$. In general, $\hat{f}_n(\cdot, \theta)$ needs to converge at a faster rate when m is less smooth.[11] Condition (d) imposes some smoothness condition on $f_0(\cdot, \theta)$ as a function of θ.[12] Condition

[11] Only α_2 matters as long as $\alpha_1 > 0$, although $\alpha_1 = \alpha_2$ in many applications.

[12] In both Examples 2.1 and 2.2, the notation $\partial f_0(\cdot, \theta_0)/\partial\theta$ is understood as $\partial f_0(x_1 + x_2^T\theta, \theta)/\partial\theta|_{\theta=\theta_0}$ since the first argument of f_0 also depends on θ.

(e) requires that $\hat{f}_n(\cdot,\theta)$ satisfy a stochastic equicontinuity-type restriction.[13] This condition is easily satisfied if $\hat{f}_n(\cdot,\theta)$ is continuously differentiable with respect to θ (e.g. $\hat{f}_n(\cdot,\theta)$ in Examples 2.1 and 2.2). More specifically, Assumption 3.4 (e) is satisfied if Assumption 3.4 (d) holds and $\partial\hat{f}_n(\cdot,\theta)/\partial\theta$ converges in probability to $\partial f_0(\cdot,\theta)/\partial\theta$ uniformly over both arguments.

Remark 3.1. It is worth while to compare conditions of Assumption 3.4 with similar ones in the literature, e.g. conditions of Theorem 2 of Chen, Linton, and Van Keilegom (2003) and Theorem 4.1 of Chen (2005). Condition (c) of Assumption 3.4 is comparable to condition (4.1.4)' of Theorem 4.1 of Chen (2005), which is weaker than conditions (2.3) and (2.4) of Theorem 2 of Chen, Linton, and Van Keilegom (2003). Condition (a) of Assumption 3.4 can be substantially weaker than similar ones imposed in Chen, Linton, and Van Keilegom (2003, Theorem 3) and Chen (2005, Lemma 4.2). For semiparametric quantile regression models such as Example 2.1 in this paper and Example 2 of Chen, Linton, and Van Keilegom (2003), $\alpha > d_1$ is needed to satisfy sufficient conditions of Chen, Linton, and Van Keilegom (2003, Theorem 3) and Chen (2005, Lemma 4.2). See Remark 3 (ii) of Chen, Linton, and Van Keilegom (2003). Even when $d_1 = 1$, the condition $\alpha > d_1$ can be substantially stronger than our condition that $\alpha > d_1/2$.

Assumption 3.5. *The following holds uniformly over θ in Θ_{δ_1}:*

$$\begin{aligned}
&\int_0^1 (1-s)\Big\{D_{ff}m^*(\theta,\hat{f}_s(\cdot,\theta))[\hat{f}_n(\cdot,\theta)-f_0(\cdot,\theta),\hat{f}_n(\cdot,\theta)-f_0(\cdot,\theta)]\Big\}\,ds\\
&-\int_0^1 (1-s)\Big\{D_{ff}m^*(\theta_0,\hat{f}_s(\cdot,\theta_0))[\hat{f}_n(\cdot,\theta_0)-f_0(\cdot,\theta_0),\hat{f}_n(\cdot,\theta_0)-f_0(\cdot,\theta_0)]\Big\}\,ds\\
&= o_p\left(n^{-1/2}\left\|\theta-\theta_0\right\|\right)+o_p\left(n^{-1}\right)+o_p(\|\theta-\theta_0\|^2)+\textit{terms not depending on }\theta,
\end{aligned}$$

where $\hat{f}_s(\cdot,\theta) = f_0(\cdot,\theta) + s(\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta))$.

This condition ensures that the remainder term by the Taylor series expansion is negligible. Assumption 3.5 is a high-level condition and more primitive conditions for this are given below.

Proposition 3.1. *(a) Assume that for any $\theta \in \Theta_{\delta_1}$, there exists $w(\theta, f(\cdot,\theta))$ such that*

$$D_{ff}m^*(\theta,f(\cdot,\theta))[h_1(\cdot),h_2(\cdot)] = \int w(\theta,f(\cdot,\theta))h_1(\cdot)h_2(\cdot)\,dP. \tag{3.9}$$

[13] We are grateful to Songnian Chen, who suggested this. In a previous version, we impose a condition on $\hat{f}_n(\cdot,\theta) - \hat{f}_n(\cdot,\theta_0) - [\partial f_0(\cdot,\theta_0)/\theta^T](\theta-\theta_0)$ rather than on $[\hat{f}_n(\cdot,\theta) - \hat{f}_n(\cdot,\theta_0)] - [f_0(\cdot,\theta) - f_0(\cdot,\theta_0)]$.

(b) Assume that one of the following holds:

(i) $w(\theta, f(\cdot,\theta))$ *does not depend on* θ *or* $f(\cdot,\theta)$ *and is bounded.*

(ii) $\|w(\theta, f(\cdot,\theta)) - w(\theta_0, f_0(\cdot,\theta_0))\| \leq C_w \|\theta - \theta_0\|$ *for some finite constant* C_w *and* $\sup_{\theta\in\Theta_{\delta_1}} \left\|\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)\right\|_\infty = o_p(n^{-1/4})$.

(iii) $\|w(\theta, f(\cdot,\theta)) - w(\theta_0, f_0(\cdot,\theta_0))\| \leq C_w [\|\theta - \theta_0\| + \|f(\cdot,\theta) - f_0(\cdot,\theta_0)\|_\infty]$ *for some finite constant* C_w *and* $\sup_{\theta\in\Theta_{\delta_1}} \left\|\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)\right\|_\infty = o_p(n^{-1/3})$.

(c) Assume that

$$Pr\left\{\left\|[\hat{f}_n(\cdot,\theta) - \hat{f}_n(\cdot,\theta_0)] - [f_0(\cdot,\theta) - f_0(\cdot,\theta_0)]\right\|_\infty \leq \delta \|\theta - \theta_0\|\right\} \geq 1 - \varepsilon.$$

Then Assumption 3.5 is satisfied.

The assumption (c) is the same as Assumption 3.4 (e).

Example 2.1 Continued: In view of (3.4), $w(\theta, f(\cdot,\theta))$ in (3.9) has the form

$$w(\theta, f(\cdot,\theta)) = p_{Y|X}(f(\cdot,\theta)|X).$$

Hence, conditions (a) and (b) of Proposition 3.1 are satisfied if $p_{Y|X}(\cdot|x)$ is Lipschitz continuous uniformly and $\sup_{\theta\in\Theta_{\delta_1}} \left\|\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)\right\|_\infty = o_p(n^{-1/3})$.

Examples 2.2 – 2.3 Continued: In Example 2.2, $w(\theta, f(\cdot,\theta)) = 1(X \in \mathcal{T})$ and in Example 2.3, $w(\theta, f(\cdot,\theta)) = 1(X \in \mathcal{T})D$. Thus, conditions (a) and (b) of Proposition 3.1 are trivially satisfied.

We place the following assumption to characterize the effect of the estimation of $f_0(\cdot,\theta)$. Later we discuss sufficient conditions for this higher level assumption.

Assumption 3.6. *(a) As a function of* θ, $D_f m^*(\theta, f_0(\cdot,\theta))[\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)]$ *is twice continuously differentiable on* Θ_{δ_1} *with probability approaching one.*

(b) There exists a d_θ*-row-vector-valued* $\Gamma_1(z)$ *such that* $E[\Gamma_1(Z)] = 0$, $E[\Gamma_1(Z)\Gamma_1^T(Z)] < \infty$ *and nonsingular,*

(3.10)
$$\frac{d}{d\theta^T}\left(D_f m^*(\theta, f_0(\cdot,\theta))[\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)]\right)\Big|_{\theta=\theta_0} = n^{-1}\sum_{i=1}^n \Gamma_1(Z_i) + o_p(n^{-1/2}).$$

The term $\Gamma_1(z)$ captures effects of first-stage nonparametric estimation of $f_0(\cdot,\theta)$. There are at least two cases in which it is easy to compute the derivatives of $D_f m^*(\theta, f_0(\cdot,\theta))[\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)]$. The first case is when $f_0(\cdot,\theta)$ does not depend on θ and the second case is when $D_f m^*(\theta, f_0(\cdot,\theta))[\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)]$ is identically zero. In Examples 2.1 and 2.2, $D_f m^*(\theta, f_0(\cdot,\theta))[\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)] = 0$ for any θ, which will be shown below. Hence, Assumption 3.6 is trivially satisfied with $\Gamma_1(z) \equiv 0$. When $D_f m^*(\theta, f_0(\cdot,\theta))[\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)] = 0$ for all $\theta \in \Theta_{\delta_1}$, no adjustment term is needed in the asymptotic distribution of $\hat{\theta}_n$.

Example 2.1 Continued: Notice that by evaluating (3.3) at $(\theta, f) = (\theta, f_0(\cdot,\theta))$:

(3.11)
$$D_f m^*(\theta, f_0(\cdot,\theta))[h] = -E[1(X \in \mathcal{T})\{\tau - 1(Y - f_0(X_1 + X_2^T\theta, \theta) \leq 0)\}h(X_1 + X_2^T\theta)] = 0,$$

where the last equality follows from the fact that $f_0(X_1 + X_2^T\theta, \theta)$ is the quantile of Y conditional on $X_1 + X_2^T\theta$ and the event that $X \in \mathcal{T}$.

Example 2.2 Continued: Suppose that h is a function of the index $x_1 + x_2\theta$. Notice that since $f_0(t,\theta)$ is the expectation of Y conditional on $X_1 + X_2^T\theta = t$ and the event that $X \in \mathcal{T}$ for each θ, the law of iterative expectations implies that in view of (3.5), $D_f m^*(\theta, f_0(X_1 + X_2^T\theta, \theta))[h(X_1 + X_2^T\theta)] \equiv 0$ for any fixed θ. Assumption 3.6 is satisfied with $\Gamma_1 \equiv 0$ for the SLS estimator whether or not the model is correctly specified. This implies that even under model misspecification, the asymptotic distribution of the SLS estimator is the same as if $f_0(\cdot,\theta)$ were known. As we discuss later, however, the asymptotic distribution under misspecification is different from that under correct specification.

We now provide sufficient conditions for Assumption 3.6 for the case when $\Gamma_1 \neq 0$. Example 2.3 is such a case. In particular, we will give an explicit expression for Γ_1 in (3.10) when $\hat{f}_n(\cdot,\theta)$ is a smooth function of θ. This case includes nonparametric kernel estimators of probability density functions and conditional expectations, as leading examples.

Let $\mathcal{L}^2(P)$ denote the L^2 space defined on the probability space of Z.

Proposition 3.2. *Assume that*

(a)

$$D_f m^*(\theta, f_0(\cdot,\theta))[h(\cdot)] = \int h(\cdot)g(\cdot,\theta)dP, \tag{3.12}$$

(b) $g(\cdot,\theta)$ is twice continuously differentiable with respect to θ with probability one,

(c) $\hat{f}_n(\cdot,\theta)$ *has an asymptotic linear form: for any* $\theta \in \Theta_{\delta_1}$,

$$\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta) = n^{-1}\sum_{j=1}^{n} \varphi_{nj}(\cdot,\theta) + b_n(\cdot,\theta) + R_n(\cdot,\theta), \tag{3.13}$$

where $\varphi_{nj}(\cdot,\theta)$ *is a stochastic term that has expectation zero (with respect to the* j*-th observation),* $b_n(\cdot,\theta)$ *is a bias term satisfying* $\sup_{z,\theta} \|b_n(z,\theta)\| = o(n^{-1/2})$ *, and* $R_n(\cdot,\theta)$ *is a remainder term satisfying* $\sup_{z,\theta} \|R_n(z,\theta)\| = o_p(n^{-1/2})$.

(d) $\hat{f}_n(\cdot,\theta)$ *is twice continuously differentiable with respect to* θ *with probability approaching one and* $\partial \hat{f}_n(\cdot,\theta)\partial\theta$ *also has an asymptotic linear form:*

$$\frac{\partial \hat{f}_n(\cdot,\theta)}{\partial\theta} - \frac{\partial f_0(\cdot,\theta)}{\partial\theta} = n^{-1}\sum_{j=1}^{n} \tilde{\varphi}_{nj}(\cdot,\theta) + o_p(n^{-1/2}), \tag{3.14}$$

uniformly over (z,θ)*, where* $\tilde{\varphi}_{nj}(\cdot,\theta)$ *is a stochastic term that has expectation zero (with respect to the* j*-th observation), and*

(e) there exists a d_θ*-row-vector-valued* $\Gamma_1(z)$ *such that* $E[\Gamma_1(Z)] = 0$ *and*

$$\max_{1\le i\le n} \|\Gamma_{n1}(Z_i) - \Gamma_1(Z_i)\| = o_p\left(n^{-1/2}\right),$$

where

$$\Gamma_{n1}(Z_i) = \int \tilde{\varphi}_{ni}(\cdot,\theta_0) g(\cdot,\theta_0)\, dP + \int \varphi_{ni}(\cdot,\theta_0)\frac{\partial g(\cdot,\theta_0)}{\partial\theta}\, dP. \tag{3.15}$$

Then Assumption 3.6 is satisfied.

Notice that under Assumption 3.2,

$$D_f m^*(\theta, f_0(\cdot,\theta))[f(\cdot,\theta) - f_0(\cdot,\theta)] = E[\Delta_2(Z, f(\cdot,\theta) - f_0(\cdot,\theta))]. \tag{3.16}$$

Thus for many cases, an expression for $g(\cdot,\theta)$ can be obtained in a straightforward manner by inspecting the form of the expectation on the right hand side of (3.16).

When $\hat{f}_n(\cdot,\theta)$ does not depend on θ, then condition (d) is trivially satisfied and the leading term has the term

$$\Gamma_{n1}(Z_i) = \int \varphi_{ni}(\cdot)\frac{\partial g(\cdot,\theta_0)}{\partial\theta}\, dP, \tag{3.17}$$

where the integral is taken with respect to the arguments of φ_{ni} and g. Example 2.3 belongs to this case.

Remark 3.2. Define $f_0(\nu(z),\theta) = E_{\mathcal{T}}[\psi(Z,\theta)|\nu(Z) = \nu(z)]$, where $\psi(z,\theta)$ is a known function of z and θ and ν is a known, d_1-vector-valued function of z. We now provide an explicit form of $\Gamma_{n1}(Z_i)$ in (3.15) when the first-stage estimator is a kernel regression estimator of $f_0(\nu(z),\theta)$ with a trimming function $\mathcal{T}_n$ and $m(z,\theta,f(\cdot,\theta))$ depends on $f(\cdot,\theta)$ only through its value $f(\nu(z),\theta)$.

Under some standard regularity conditions, $\varphi_{ni}(\cdot,\theta)$ and $\tilde{\varphi}_{ni}(\cdot,\theta)$ in (3.13) and (3.14) have the form:

(3.18)

$$\varphi_{ni}(\nu(z),\theta) = n^{-1}\sum_{i=1}^{n} 1(\nu(X_i)\in\mathcal{T})\frac{\psi(Z_i,\theta) - E_{\mathcal{T}}[\psi(Z,\theta)|\nu(Z)=\nu(Z_i)]}{h_n^{d_1} p_{\mathcal{T}}(\nu(z))} K\left(\frac{\nu(z)-\nu(Z_i)}{h_n}\right),$$

and

$$\tilde{\varphi}_{ni}(\nu(z),\theta) = n^{-1}\sum_{i=1}^{n} 1(\nu(X_i)\in\mathcal{T})\frac{\partial\psi(Z_i,\theta)/\partial\theta - E_{\mathcal{T}}[\partial\psi(Z,\theta)/\partial\theta|\nu(Z)=\nu(Z_i)]}{h_n^{d_1} p_{\mathcal{T}}(\nu(z))} K\left(\frac{\nu(z)-\nu(Z_i)}{h_n}\right),$$

where $p_{\mathcal{T}}(\nu)$ is the joint density of $\nu(Z)$ and $1(\nu(Z)\in\mathcal{T})$. Then by usual changes of variables,

(3.19)

$$\begin{aligned}\Gamma_1(Z_i) = {} & \{\partial\psi(Z_i,\theta_0)/\partial\theta - E_{\mathcal{T}}[\partial\psi(Z,\theta_0)/\partial\theta|\nu(Z)=\nu(Z_i)]\}\, E_{\mathcal{T}}\left[g(Z,\theta_0)|\nu(X)=\nu(X_i)\right] \\ & + \{\psi(Z_i,\theta_0) - E_{\mathcal{T}}[\psi(Z,\theta_0)|\nu(Z)=\nu(Z_i)]\}\, E_{\mathcal{T}}\left[\frac{\partial g(Z,\theta_0)}{\partial\theta}\middle|\nu(X)=\nu(X_i)\right].\end{aligned}$$

Although we have only worked out details for the case of the kernel mean regression estimator, it is straightforward to develop analogous results for other kernel-type estimators.

Example 2.3 Continued: It follows from (3.6) that

$$g(z,\theta) = 1(x\in\mathcal{T})d\{\theta - (y_1 - f_0(x))\}.$$

Also, $E_{\mathcal{T}}[\psi(Z,\theta)|\nu(Z)=\nu(z)] = E_{\mathcal{T}}[Y_0|D=0,X=x]$ and thus, by (3.19),

$$\Gamma_1(Z_i) = 1(X_i\in\mathcal{T})(1-D_i)[Y_{0i} - E_{\mathcal{T}}[Y_0|D=0,X=x]]\frac{p_{\mathcal{T}}(X_i,D_i=1)}{p_{\mathcal{T}}(X_i,D_i=0)}. \tag{3.20}$$

Note that $p_{\mathcal{T}}(X_i,D=1)/p_{\mathcal{T}}(X_i,D=0)$ appears in the expression of $\Gamma_1(Z_i)$ because the first-stage estimation uses the $D=0$ sample and the second-stage estimation uses the $D=1$ sample.

3.2 Theorems

This subsection presents the main results of the paper. Let $\Delta_{10}(z)$ and $\Delta_{20}(z,h)$ denote $\Delta_1(z)$ and $\Delta_2(z,h)$ in Assumption 3.2 with $(\theta_1, f_1) = (\theta, f)$ and $(\theta_2, f_2) = (\theta_0, f_0(\cdot,\theta_0))$. Thus, $\Delta_{10}(z)(\theta-\theta_0)+\Delta_{20}(z, f(\cdot,\theta)-f_0(\cdot,\theta_0))$ is a linear approximation of $m(z,\theta,f(\cdot,\theta))-m(z,\theta_0,f_0(\cdot,\theta_0))$. Define $\Delta^*_{20}[h] = E[\Delta_{20}(Z,h)]$ for fixed h. Also define a d_θ-row-vector-valued function $\Gamma_0(z)$ such that

$$\Gamma_0(z) = \Delta_{10}(z) - E[\Delta_{10}(Z)] + \Delta_{20}\left[z, \frac{\partial f_0(\cdot,\theta_0)}{\partial\theta^T}\right] - \Delta^*_{20}\left[\frac{\partial f_0(\cdot,\theta_0)}{\partial\theta^T}\right] + \Gamma_1(z), \tag{3.21}$$

$\Omega_0 = E[\Gamma_0(Z)^T\Gamma_0(Z)]$, and

$$V_0 = \frac{d^2\, m^*(\theta, f_0(\cdot,\theta))}{d\theta\, d\theta^T}\Big|_{\theta=\theta_0}.$$

Notice that V_0 is the Hessian matrix of $m^*(\theta, f_0(\cdot,\theta))$ with respect to θ, evaluated at $\theta=\theta_0$. The following theorem gives the asymptotic distribution of $\hat\theta_n$.

Theorem 3.3. *Assume that $\{Z_i : i = 1,\dots,n\}$ are a random sample of Z. Let Assumptions 3.1-3.6 hold. Assume that there exists $C(z)$ satisfying $\|\Delta_{20}[z,h(\cdot,\theta)]\| \le C(z)\,\|h(\cdot,\theta)\|_\infty$ for any θ and $\|C(Z)\|_{L^2(P)} < \infty$. Also, assume that Ω_0 exists and V_0 is a positive definite matrix. Then*

$$n^{1/2}(\hat\theta_n - \theta_0) \to_d \mathbf{N}(0, V_0^{-1}\Omega_0 V_0^{-1}).$$

Let $\partial_1 m^*(\theta,f)$ denote a vector of the usual partial derivatives of $m^*(\theta,f)$ with respect to the first argument θ. In this notation, $\partial_1 m^*(\theta, f(\cdot,\theta))$ denotes the partial derivative of $m^*(\theta,f)$ with respect to the first argument θ, evaluated at $(\theta,f) = (\theta, f(\cdot,\theta))$. Similarly, let $\partial_1^2 m^*(\theta,f)$ denote the usual Hessian matrix of $m^*(\theta,f)$ with respect to θ, holding f constant. Using this notation, note that by the chain rule, the expression of V_0 can be written as[14]

$$\begin{aligned} V_0 &= \frac{d^2\, m^*(\theta, f_0(\cdot,\theta))}{d\theta\, d\theta^T}\Big|_{\theta=\theta_0} \\ &= \partial_1^2 m^*(\theta_0, f_0(\cdot,\theta_0)) + D_{ff} m^*(\theta_0, f_0(\cdot,\theta_0))\left[\frac{\partial f_0(\cdot,\theta_0)}{\partial\theta}, \frac{\partial f_0(\cdot,\theta_0)}{\partial\theta^T}\right] \\ &\quad + 2\left\{D_f\left[\partial_1 m^*(\theta_0, f_0(\cdot,\theta_0))^T\right]\left[\frac{\partial f_0(\cdot,\theta_0)}{\partial\theta}\right]\right\} + D_f m^*(\theta_0, f_0(\cdot,\theta_0))\left[\frac{\partial^2 f_0(\cdot,\theta_0)}{\partial\theta\partial\theta^T}\right]. \end{aligned} \tag{3.22}$$

[14] See Appendix A for the expression of V_0 when $d_f > 1$.

We now modify the main theorem for an important special case when $f_0(\cdot,\theta)$ is not a function of θ, i.e. $f_0(\cdot,\theta) \equiv f_0(\cdot)$. In this case, the objective function can be less smooth with respect to the nonparametric part in Assumption 3.2 and Assumption 3.4 can be weakened in an obvious manner. Define $\mathcal{F}_{\delta_2} = \{f \in \mathcal{F} : \|f(\cdot) - f_0(\cdot)\|_\infty < \delta_2\}$.

Assumption 3.7. *For any (θ_1, f) and (θ_2, f) in $\Theta_{\delta_1} \times \mathcal{F}_{\delta_2}$, there exist a d_θ-row-vector-valued function $\Delta_1(z, \theta_2, f)$ and a function $\dot{m}(z, \delta_1)$ satisfying*

$$(a) \quad |m(z,\theta_1,f(\cdot)) - m(z,\theta_2,f(\cdot)) - \Delta_1(z,\theta_2,f)(\theta_1-\theta_2)| \le \|\theta_1-\theta_2\|\,\dot{m}(z,\delta_1),$$

$$(b) \quad \|\dot{m}(Z,\delta_1)\|_{L^2(P)} \le C\delta_1^{\alpha_1} \quad \textit{for some constants } C<\infty \textit{ and } \alpha_1>0,$$

and

$$(c) \quad \sup_{f\in\mathcal{F}_{\delta_2}} \left\| n^{-1}\sum_{i=1}^n \{\Delta_1(Z_i,\theta_0,f) - E[\Delta_1(Z,\theta_0,f)]\} - \{\Delta_1(Z_i,\theta_0,f_0) - E[\Delta_1(Z,\theta_0,f_0)]\} \right\| = o_p\left(n^{-1/2}\right) \quad \textit{for any } \delta_2 \to 0.$$

Note that by conditions (a) and (b), m is assumed to have a linear expansion with respect to only θ along with a restriction on the remainder term. Condition (c) is a high-level, stochastic equicontinuity condition that can be verified, for example, using Sections 4 and 5 of Andrews (1994b) and Section 4 of Chen, Linton, and Van Keilegom (2003). In particular, Chen, Linton, and Van Keilegom (2003, Theorem 3) distinguish the case when $\Delta_1(z,\theta_0,f)$ is pointwise continuous from the case when $\Delta_1(z,\theta_0,f)$ is not.

When $\Delta_1(z,\theta_0,f)$ is not pointwise continuous with respect to h, then there exists an interesting tradeoff between Assumption 3.2 and Assumption 3.7. In this case, to use Assumption 3.7, it may be necessary to assume a smaller function space for $f_0(\cdot)$ (e.g., $\mathcal{C}_M^\alpha(\mathcal{X})$ with $\alpha > d_1$ rather than $\alpha > d_1/2$) to verify the stochastic equicontinuity condition (see (3.2) of Theorem 3 of Chen, Linton, and Van Keilegom (2003)), whereas conditions (a) and (b) of Assumption 3.7 are weaker than Assumption 3.2.[15]

Assumption 3.8. *(a) $f_0(\cdot)$ is an element of $\mathcal{C}_M^\alpha(\mathcal{X})$ for some $\alpha > d_1/2$, where d_1 is the dimension of the argument of $f_0(\cdot)$ and $\mathcal{X}$ is a finite union of bounded, convex subset of $\mathbf{R}^{d_1}$ with nonempty interior.*

(b) $\hat{f}_n(\cdot) \in \mathcal{C}_M^\alpha(\mathcal{X})$ with probability approaching one.

[15] In this respect, it appears that it is better to use Assumption 3.2 than Assumption 3.7 when both assumptions are satisfied. However, there are cases for which only Assumption 3.7 is satisfied. See, e.g., an estimator of hit rates in Chen, Linton, and Van Keilegom (2003, Example 1).

(c) $\left\| \hat{f}_n(\cdot) - f_0(\cdot) \right\|_\infty = o_p(1)$.

The following theorem gives the asymptotic distribution of $\hat{\theta}_n$ when the first-stage nonparametric estimator $\hat{f}_n(\cdot, \theta)$ does not depend on θ.

Theorem 3.4. *Assume that* $\{Z_i : i = 1, \ldots, n\}$ *are a random sample of* Z*. Let Assumptions 3.1, 3.3, 3.5, 3.6, and 3.8 hold. Assume that either Assumption 3.2 or Assumption 3.7 holds. Also, assume that* $\Omega_0 = E[\Gamma_0(Z)^T \Gamma_0(Z)^T]$ *exists and* V_0 *is a positive definite matrix, where*

$$\Gamma_0(z) = \Delta_1(z, \theta_0, f_0) - E\left[\Delta_1(Z, \theta_0, f_0)\right] + \Gamma_1(z)$$

and

$$V_0 = \frac{\partial^2 m^*(\theta_0, f_0(\cdot))}{\partial \theta \partial \theta^T}.$$

Then

$$n^{1/2}(\hat{\theta}_n - \theta_0) \to_d \mathbf{N}(0, V_0^{-1} \Omega_0 V_0^{-1}).$$

3.3 Analysis of Effects of the First-Stage Estimation

This section provides some analysis of the correction term $\Gamma_1(z)$ in (3.10). We begin with a sufficient condition under which the first-stage nonparametric estimation does not affect the asymptotic distribution of $\hat{\theta}_n$. This is called an asymptotic orthogonality condition between θ_0 and f_0 (Andrews (1994a), equation 2.12). Newey (1994) discusses conditions for the asymptotic orthogonality (see Propositions 2 and 3 of Newey (1994)).

To describe an asymptotic orthogonality condition in our setup, define $\mathcal{H} = \{h(\cdot) \in \mathcal{F} : \mathbf{R}^{d_1} \mapsto \mathbf{R}^{d_f}\}$, that is a subset of $\mathcal{F}$ such that an element of $\mathcal{H}$ has the same arguments as $f_0(\cdot, \theta)$ for each θ.

Theorem 3.5. *If* $D_f m^*(\theta, f_0(\cdot, \theta))[h(\cdot)] = 0$ *for any* $\theta \in \Theta_{\delta_1}$ *and for any* $h(\cdot) \in \mathcal{H}$*, then* $\Gamma_1(z) \equiv 0$*. That is,* $\hat{\theta}_n$ *has the same asymptotic distribution that it would have if* $f_0(\cdot, \theta)$ *were known.*

As shown already in Section 3.1, the assumption of Theorem 3.5 is satisfied in Examples 2.1 and 2.2. There are a number of examples in which this assumption is not satisfied, including Example 2.3, sample selection models with a nonparametric selection mechanism (e.g., Ahn and Powell (1993) and Das, Newey, and Vella (2003)), average derivative estimators (e.g., Powell, Stock, and Stoker (1989)), and regression estimators with generated regressors (e.g., Ahn and Manski (1993)).

It is interesting to see the connection between Theorem 3.5 and Propositions 2 and 3 of Newey (1994). Theorem 3.5 can be viewed as an analogous version of Proposition 2 of Newey (1994) for non-smooth semiparametric estimators. Furthermore, as in Examples 2.1 and 2.2, the assumption of Theorem 3.5 can be verified using the law of iterative expectations, which is reminiscent of Proposition 3 of Newey (1994). In this regard, Theorem 3.5 provides a unifying interpretation of two apparently different results of Newey (1994, Propositions 2 and 3).

To understand the effects of first-stage estimation more carefully, notice that by simple calculus, the left-hand side of (3.10) can be written as

$$
\begin{aligned}
&\frac{d}{d\theta^T}\left(D_f m^*(\theta, f_0(\cdot,\theta))[\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)]\right)\Big|_{\theta=\theta_0} \\
&= D_f m^*(\theta_0, f_0(\cdot,\theta_0))\left[\frac{\partial \hat{f}_n(\cdot,\theta_0)}{\partial \theta^T} - \frac{\partial f_0(\cdot,\theta_0)}{\partial \theta^T}\right] \\
&+ \left\{D_f\left[\partial_1 m^*(\theta_0, f_0(\cdot,\theta_0))\right]\left[\hat{f}_n(\cdot,\theta_0) - f_0(\cdot,\theta_0)\right]\right\}^T \\
&+ D_{ff} m^*(\theta_0, f_0(\cdot,\theta_0))\left[\hat{f}_n(\cdot,\theta_0) - f_0(\cdot,\theta_0), \frac{\partial f_0(\cdot,\theta_0)}{\partial \theta^T}\right],
\end{aligned}
\tag{3.23}
$$

where the first and third terms appear because both $\hat{f}_n(\cdot,\theta)$ and $f_0(\cdot,\theta)$ may depend on θ and the second term shows up because of possible interactions between θ and f in the definition of $m^*(\theta, f)$. In Example 2.3, only the second term of the right-hand side of (3.23) is non-zero since $f_0(\cdot,\theta)$ does not depend on θ. If the first term of the right-hand side of (3.23) is non-zero and is not cancelled out by other terms, then that is the case when $\partial \hat{f}_n(\cdot,\theta_0)/\partial\theta$ affects the asymptotic distribution; however, all the existing estimators we have examined do not belong to this case.

4 Examples

This section gives asymptotic distributions of M-estimators considered in Examples 2.1 – 2.3.

4.1 Single-Index Quantile Regression Models

For simplicity, assume that $P_{Y|X}(y|x) \equiv P_{Y|X_1+X_2^T\theta_0}(y|x_1 + x_2'\theta_0)$, that is the conditional distribution of Y given X depends only on the index $x_1 + x_2'\theta_0$. Let $p_{U|X_1+X_2^T\theta_0}(0|t)$ be the PDF of U conditional on $X_1 + X_2^T\theta_0 = t$, and $\dot{P}_{U|X_1+X_2^T\theta_0}[0|t]$ the partial derivative of $P_{U|X_1+X_2^T\theta_0}[0|t]$ with respect to t. The weak consistency of $\hat{\theta}_n$ to θ_0 is given in the Appendix

(see Lemma B.8). Using arguments similar to those used in Klein and Spady (1993, pp. 401-403) and also the Implicit Function Theorem, it is not difficult to show that

(4.1)
$$\frac{\partial G_0(x_1 + x_2^T\theta_0, \theta_0)}{\partial \theta} = \frac{\dot{P}_{U|X_1+X_2^T\theta_0}(0|x_1 + x_2'\theta_0)}{p_{U|X_1+X_2^T\theta_0}(0|x_1 + x_2'\theta_0)} \left(x_2 - E[X_2|X_1 + X_2'\theta_0 = x_1 + x_2^T\theta_0, X \in \mathcal{T}]\right).$$

Then by Theorem 3.3,

$$n^{1/2}(\hat{\theta}_n - \theta_0) \to_d \mathbf{N}(0, V_0^{-1}\Omega_0 V_0^{-1}),$$

where

$$\Omega_0 = \tau(1-\tau)E\left[1(X \in \mathcal{T})\frac{\partial G_0(X_1 + X_2^T\theta_0, \theta_0)}{\partial \theta}\frac{\partial G_0(X_1 + X_2^T\theta_0, \theta_0)}{\partial \theta^T}\right]$$

and

$$V_0 = E\left[1(X \in \mathcal{T})p_{U|X_1+X_2^T\theta_0}(0|X_1 + X_2^T\theta_0)\frac{\partial G_0(X_1 + X_2^T\theta_0, \theta_0)}{\partial \theta}\frac{\partial G_0(X_1 + X_2^T\theta_0, \theta_0)}{\partial \theta^T}\right].$$

The asymptotic variance can be estimated consistently by a sample analog estimator based on the expressions of Ω_0, V_0, and (4.1).

4.2 Semiparametric Least Squares Estimation under Misspecification

The asymptotic distribution of the SLS estimator is established by Ichimura (1993) under the assumption that the model is correctly specified, that is $E[Y|X = x] = f_0(x_1 + x_2^T\theta_0, \theta_0)$. In this section, we establish the asymptotic distribution of the SLS estimator when $E[Y|X = x]$ may not belong to a class of single-index models.

It follows from (3.21) that

$$\Omega_0 = E\left[1(X \in \mathcal{T})\{Y - f_0(X_1 + X_2^T\theta_0, \theta_0)\}^2\frac{\partial f_0(X_1 + X_2^T\theta_0, \theta_0)}{\partial \theta}\frac{\partial f_0(X_1 + X_2^T\theta_0, \theta_0)}{\partial \theta^T}\right].$$

In addition, observe that by (3.22),

(4.2)
$$\begin{aligned} V_0 &= D_{ff}m^*(\theta_0, f_0(\cdot, \theta_0))\left[\frac{\partial f_0(\cdot, \theta_0)}{\partial \theta}, \frac{\partial f_0(\cdot, \theta_0)}{\partial \theta^T}\right] \\ &\quad + D_f m^*(\theta_0, f_0(\cdot, \theta_0))\left[\frac{\partial^2 f_0(\cdot, \theta_0)}{\partial \theta \partial \theta^T}\right] \\ &= E\left[1(X \in \mathcal{T})\frac{\partial f_0(X_1 + X_2^T\theta_0, \theta_0)}{\partial \theta}\frac{\partial f_0(X_1 + X_2^T\theta_0, \theta_0)}{\partial \theta^T}\right] \\ &\quad - E\left[1(X \in \mathcal{T})\{Y - f_0(X_1 + X_2^T\theta_0, \theta_0)\}\frac{\partial^2 f_0(X_1 + X_2^T\theta_0, \theta_0)}{\partial \theta \partial \theta^T}\right]. \end{aligned}$$

Notice that the second term in the expression of V_0 is zero only when the model is correctly specified. Then by Theorem 3.3 combined with results obtained in this section, we have, under model misspecification,

$$n^{1/2}(\hat{\theta}_n - \theta_0) \to_d \mathbf{N}\left(0, V_0^{-1}\Omega_0 V_0^{-1}\right). \tag{4.3}$$

The asymptotic variance in (4.3) is different from the asymptotic variance when the model is correctly specified.

This suggests a new asymptotic variance estimator $\hat{V}_n^{-1}\hat{\Omega}_n\hat{V}_n^{-1}$, where

$$\hat{\Omega}_n = n^{-1}\sum_{i=1}^n 1(X_i \in \mathcal{T})\{Y_i - \hat{f}_n(X_{1i} + X_{2i}^T\hat{\theta}_n, \hat{\theta}_n)\}^2 \frac{\partial \hat{f}_n(X_{1i} + X_{2i}^T\hat{\theta}_n, \hat{\theta}_n)}{\partial \theta} \frac{\hat{f}_n(X_{1i} + X_{2i}^T\hat{\theta}_n, \hat{\theta}_n)}{\partial \theta^T}$$

and

$$\begin{aligned}\hat{V}_n = {} & n^{-1}\sum_{i=1}^n 1(X_i \in \mathcal{T})\frac{\hat{f}_n(X_{1i} + X_{2i}^T\hat{\theta}_n, \hat{\theta}_n)}{\partial \theta} \frac{\hat{f}_n(X_{1i} + X_{2i}^T\hat{\theta}_n, \hat{\theta}_n)}{\partial \theta^T} \\ & - n^{-1}\sum_{i=1}^n 1(X_i \in \mathcal{T})\{Y_i - \hat{f}_n(X_{1i} + X_{2i}^T\hat{\theta}_n, \hat{\theta}_n)\}\frac{\partial^2 \hat{f}_n(X_{1i} + X_{2i}^T\hat{\theta}_n, \hat{\theta}_n)}{\partial \theta \partial \theta^T}.\end{aligned}$$

In contrast to a sample analog estimator of the asymptotic variance of the SLS estimator of Ichimura (1993, Theorem 7.1), the new asymptotic variance estimator is consistent whether or not the model is correctly specified. The result in this section can be viewed as a semiparametric analog of Theorem 3.3 and Corollary 3.4 of White (1981), who characterizes the asymptotic distribution of parametric least squares for misspecified nonlinear regression models.

4.3 Smoothed Matching Estimator

Note that $\Delta_{10}(z) = 1(x \in \mathcal{T})d[\theta - (y_1 - f_0(x))]$. It follows from (3.20) that

$$\begin{aligned}\Gamma_0(z) = {} & 1(x \in \mathcal{T})d[\theta - (y_1 - f_0(x))] \\ & + 1(x \in \mathcal{T})(1-d)[y_0 - f_0(x)]\frac{p_{\mathcal{T}}(x, D=1)}{p_{\mathcal{T}}(x, D=0)}.\end{aligned} \tag{4.4}$$

Then

$$\begin{aligned}\Omega_0 = {} & E\left[\{\theta_0 - (Y_1 - f_0(X))\}^2 \middle| D = 1, X \in \mathcal{T}\right]\Pr(D=1, X \in \mathcal{T}) \\ & + E\left[\mathrm{Var}(Y_0|X, D=0, X \in \mathcal{T})\frac{p_{\mathcal{T}}^2(X|D=1)}{p_{\mathcal{T}}^2(X|D=0)}\right]\frac{\Pr^2(D=1, X \in \mathcal{T})}{\Pr(D=0, X \in \mathcal{T})}.\end{aligned}$$

Note that $V_0 = \Pr(D = 1, X \in \mathcal{T})$. Then by Theorem 3.3,

$$n^{1/2}(\hat{\theta}_n - \theta_0) \to_d \mathbf{N}(0, \Sigma_0),$$

where

$$\begin{aligned}\Sigma_0 = {} & E\left[\{\theta_0 - (Y_1 - f_0(X))\}^2 \middle| D = 1, X \in \mathcal{T}\right] [\Pr(D = 1, X \in \mathcal{T})]^{-1} \\ & + E\left[\mathrm{Var}(Y_0|X, D = 0, X \in \mathcal{T}) \frac{p_{\mathcal{T}}^2(X|D = 1)}{p_{\mathcal{T}}^2(X|D = 0)}\right] [\Pr(D = 0, X \in \mathcal{T})]^{-1}.\end{aligned}$$

This result corresponds to the result of Heckman, Ichimura, and Todd (1998) except that they make a choice-based sampling assumption (independent and identically distributed within each group) and we make a random sampling assumption on (Y, X, D).

Appendix

A Theorem for the General Case

It is straightforward to extend Theorem 3.3 for the general case. When $d_f > 1$, the asymptotic variance has the same form $V_0^{-1} E[\Gamma_0(Z)^T \Gamma_0(Z)] V_0^{-1}$ with general forms of V_0 and $\Gamma_0(z)$:

(A.1)

$$\begin{aligned}V_0 = {} & \frac{\partial^2 m^*(\theta_0, \mathbf{f_0}(\cdot, \theta_0))}{\partial\theta\partial\theta^T} + \sum_{j=1}^{d_f}\sum_{k=1}^{d_f} D_{f_j f_k} m^*(\theta_0, \mathbf{f_0}(\cdot, \theta_0)) \left[\frac{\partial f_{0j}(\cdot, \theta_0)}{\partial\theta}, \frac{\partial f_{0k}(\cdot, \theta_0)}{\partial\theta^T}\right] \\ & + 2\left\{\sum_{j=1}^{d_f} D_{f_j}\left[\partial_1 m^*(\theta_0, \mathbf{f_0}(\cdot, \theta_0))^T\right]\left[\frac{\partial f_{0j}(\cdot, \theta_0)}{\partial\theta}\right]\right\} + \sum_{j=1}^{d_f} D_{f_j} m^*(\theta_0, \mathbf{f_0}(\cdot, \theta_0))\left[\frac{\partial^2 f_{0j}(\cdot, \theta_0)}{\partial\theta\partial\theta^T}\right]\end{aligned}$$

and

(A.2)

$$\Gamma_0(z) = \Delta_{10}(z) - E[\Delta_{10}(Z)] + \sum_{j=1}^{d_f}\left\{\Delta_{20}\left[z, \frac{\partial f_{0j}(\cdot, \theta_0)}{\partial\theta^T}\right] - \Delta_{20}^*\left[\frac{\partial f_{0j}(\cdot, \theta_0)}{\partial\theta^T}\right]\right\} + \mathbf{\Gamma_1}(z),$$

where $\mathbf{f_0}(\cdot,\theta_0) = [f_{01}(\cdot,\theta_0),\ldots,f_{0d_f}(\cdot,\theta_0)]$, $\hat{\mathbf{f}}_\mathbf{n}(\cdot,\theta_0) = [\hat{f}_{n1}(\cdot,\theta_0),\ldots,\hat{f}_{nd_f}(\cdot,\theta_0)]$ and $\mathbf{\Gamma_1}$ is the leading term of the asymptotic expansion of $\frac{d}{d\theta^T}\left(D_f m^*(\theta,\mathbf{f_0}(\cdot,\theta))[\hat{\mathbf{f}}_\mathbf{n}(\cdot,\theta)-\mathbf{f_0}(\cdot,\theta)]\right)\Big|_{\theta=\theta_0}$:

$$
\begin{aligned}
&\frac{d}{d\theta^T}\left(D_f m^*(\theta,\mathbf{f_0}(\cdot,\theta))[\hat{\mathbf{f}}_\mathbf{n}(\cdot,\theta)-\mathbf{f_0}(\cdot,\theta)]\right)\Big|_{\theta=\theta_0}\\
&=\sum_{j=1}^{d_f} D_{f_j} m^*(\theta_0,\mathbf{f_0}(\cdot,\theta_0))\left[\frac{\partial \hat{f}_{nj}(\cdot,\theta_0)}{\partial\theta^T}-\frac{\partial f_{0j}(\cdot,\theta_0)}{\partial\theta^T}\right]\\
&+\left\{\sum_{j=1}^{d_f} D_{f_j}\left[\partial_1 m^*(\theta_0,\mathbf{f_0}(\cdot,\theta_0))\right]\left[\hat{f}_{nj}(\cdot,\theta_0)-f_{0j}(\cdot,\theta_0)\right]\right\}^T\\
&+\sum_{j=1}^{d_f}\sum_{k=1}^{d_f} D_{f_j f_k} m^*(\theta_0,\mathbf{f_0}(\cdot,\theta_0))\left[\hat{f}_{nj}(\cdot,\theta_0)-f_{0j}(\cdot,\theta_0),\frac{\partial f_{0k}(\cdot,\theta_0)}{\partial\theta^T}\right].
\end{aligned}
\tag{A.3}
$$

B Proofs

Throughout the proofs, we will use $C>0$ to denote a generic finite constant that may be different in different uses. When it is necessary to denote a particular constant, then we will use a C with a subscript.

Proof of Theorem 3.3. To prove the theorem, define

$$R(z,\theta,f) = m(z,\theta,f(\cdot,\theta)) - m(z,\theta_0,f_0(\cdot,\theta_0)) - \Delta_{10}(z)(\theta-\theta_0) - \Delta_{20}[z,f(\cdot,\theta)-f_0(\cdot,\theta_0)].$$

As shorthand notation, let $m_i(\theta,f) = m(Z_i,\theta,f(\cdot,\theta))$, $m_i(\theta_0,f_0(\cdot,\theta_0)) = m(Z_i,\theta_0,f(\cdot,\theta_0))$, $\Delta_{1i}(\theta-\theta_0) = \Delta_{10}(Z_i)(\theta-\theta_0)$, $\Delta_{2i}(f(\cdot,\theta)-f_0(\cdot,\theta_0)) = \Delta_{20}(Z_i,f(\cdot,\theta)-f_0(\cdot,\theta_0))$, and $R_i(\theta,f) = R(Z_i,\theta,f)$. Define

$$S_n(\theta,f) = n^{-1}\sum_{i=1}^n \left[m_i(\theta,f) - m_i(\theta_0,f_0(\cdot,\theta_0))\right].$$

Also, let $R^*(\theta,f) = E[R_i(\theta,f)]$ for fixed θ and f. Recall that $\Delta_{20}^*[h] = E[\Delta_{20}(Z,h)]$ for fixed h.

Write

$$S_n(\theta,f) = S_{n1}(\theta) + S_{n2}(f) + S_{n3}(\theta,f) + S^*(\theta,f),$$

where

$$S_{n1}(\theta) = n^{-1}\sum_{i=1}^{n}\left[\Delta_{1i} - E(\Delta_{1i})\right](\theta - \theta_0),$$
$$S_{n2}(f) = n^{-1}\sum_{i=1}^{n}\Delta_{2i}[f - f_0(\cdot,\theta_0)] - \Delta_{20}^{*}[f - f_0(\cdot,\theta_0)],$$
$$S_{n3}(\theta, f) = n^{-1}\sum_{i=1}^{n}R_i(\theta, f) - R^{*}(\theta, f), \quad \text{and}$$
$$S^{*}(\theta, f) = m^{*}(\theta, f) - m^{*}(\theta_0, f_0(\cdot,\theta_0)).$$

Notice that $\hat{\theta}_n$ minimizes $S_n(\theta, \hat{f}_n(\cdot,\theta))$ and θ_0 minimizes $S^{*}(\theta, f_0(\cdot,\theta))$. Also, recall that $\Theta_{\delta_1} = \{\theta \in \Theta : \|\theta - \theta_0\| < \delta_1\}$ and $\mathcal{F}_{\delta_1,\delta_2} = \{f \in \mathcal{F} : \sup_{\theta\in\Theta_{\delta_1}} \|f(\cdot,\theta) - f_0(\cdot,\theta_0)\|_\infty < \delta_2\}$.

Define

$$\hat{\Gamma}_n = n^{-1}\sum_{i=1}^{n}\left[\Delta_{1i} - E(\Delta_{1i})\right] + [\Delta_{2i} - \Delta_{20}^{*}][\partial f_0(\cdot,\theta_0)/\partial\theta^T]$$
$$+ \frac{d}{d\theta^T}\Big(D_f m^{*}(\theta, f_0(\cdot,\theta))[\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)]\Big)\Big|_{\theta=\theta_0}.$$

For any $\delta_1 \to 0$ and $\delta_2 \to 0$, by Lemmas B.3-B.7 in subsections B.1-B.3,

$$\begin{aligned} S_n(\theta, \hat{f}_n(\cdot,\theta)) &= \frac{1}{2}(\theta - \theta_0)^T V_0(\theta - \theta_0) + \hat{\Gamma}_n(\theta - \theta_0) \\ &\quad + O_p\left[n^{-1/2}(\delta_1 + \delta_2)\left(\delta_1^{\alpha_1} + \delta_2^{\alpha_2}\right)\right] + o_p(n^{-1/2}\delta_1) \\ &\quad + o_p\left(n^{-1}\right) + o_p\left(\|\theta - \theta_0\|^2\right) + R_S, \end{aligned} \tag{B.1}$$

uniformly over $\theta \in \Theta_{\delta_1}$, where R_S is a term that is independent of θ.

Notice that $\hat{\Gamma}_n = O_p(n^{-1/2})$ in view of (3.10). The theorem can be proved by applying Theorems 1 and 2 of Sherman (1994) to (B.1). By Theorem 1 of Sherman (1994),

$$\left\|\hat{\theta}_n - \theta_0\right\| = \max[O_p(\varepsilon_n^{1/2}) + o_p(n^{-1/4}\delta_1^{1/2}), O_p(n^{-1/2})], \tag{B.2}$$

where $\varepsilon_n = n^{-1/2}(\delta_1 + \delta_2)\left(\delta_1^{\alpha_1} + \delta_2^{\alpha_2}\right)$. As in Sherman (1994, comments following Theorem 1), we first obtain an initial rate of convergence when $\delta_1 \to 0$. Note that

$$\|f(\cdot,\theta) - f_0(\cdot,\theta_0)\|_\infty \le \|f(\cdot,\theta) - f_0(\cdot,\theta)\|_\infty + C\|\theta - \theta_0\|$$

for some constant C. Hence, when $\delta_1 \to 0$ and $\delta_2 \to 0$, (B.2) implies that $\left\|\hat{\theta}_n - \theta_0\right\| = o_p(n^{-1/4})$. Then we shrink the parameter spaces Θ_{δ_1} and $\mathcal{F}_{\delta_1,\delta_2}$ by taking δ_1 satisfying $n^{1/4}\delta_1 \to 0$ and $\delta_2 = C\max\{\tilde{\delta}_2, \delta_1\}$ with some constant C. It follow from (B.2) that the

convergence rate can be improved such that $\left\|\hat{\theta}_n - \theta_0\right\| = o_p(n^{-3/8})$. Repeated applications of (B.2) give $\left\|\hat{\theta}_n - \theta_0\right\| = O_p(n^{-1/2})$, provided that $n^{1/2}\tilde{\delta}_2^{1+\alpha_2} \to 0$. Note that $\hat{\Gamma}_n$ converges in distribution to $\mathbf{N}(0, \Omega_0)$ by (3.10) and the central limit theorem. Then the theorem follows by applying Theorem 2 of Sherman (1994) to (B.1). □

B.1 Asymptotic expansion of $S_{n3}(\theta, f)$

Let $\mathcal{H}$ be a class of measurable functions with a measurable envelope function H. Let $N(\varepsilon, \mathcal{H}, \|\cdot\|_{\mathcal{H}})$ and $N_{[]}(\varepsilon, \mathcal{H}, \|\cdot\|_{\mathcal{H}})$, respectively, denote the covering and bracketing numbers for the set $\mathcal{H}$ (for exact definitions, see, for example, Van der Vaart and Wellner (1996, p.83)). In addition, let $J_{[]}(1, \mathcal{H}, L^2(P))$ denote a bracketing integral of $\mathcal{H}$, that is

$$J_{[]}(1, \mathcal{H}, L^2(P)) = \int_0^1 \sqrt{1 + \log N_{[]}(\varepsilon \|H\|_{L^2(P)}, \mathcal{H}, L^2(P))}\, d\varepsilon.$$

We will use the following lemmas. The first lemma is due to the last display of Theorem 2.14.2 of Van der Vaart and Wellner (1996, p.240).

Lemma B.1. *Let $\mathcal{H}$ be a class of measurable functions with a measurable envelope function H. Then there exists a constant C such that*

$$E\left[\sup_{h\in\mathcal{H}}\left|n^{-1/2}\sum_{i=1}^{n}\{h(Z_i) - E[h(Z)]\}\right|\right] \leq C J_{[]}(1, \mathcal{H}, L^2(P))\, \|H\|_{L^2(P)}.$$

Lemma B.2. *Let $\mathcal{F}_1$ be a class of functions $f : \mathcal{Z} \times \Theta \mapsto \mathbf{R}^{d_f}$ such that there exists a universal constant C_L satisfying*

$$\|f(z, \theta_1) - f(z, \theta_2)\| \leq C_L \|\theta_1 - \theta_2\| \tag{B.3}$$

for any $f \in \mathcal{F}_1$. Also, assume that for each fixed $\bar{\theta} \in \Theta$, the subclass $\{f(z, \bar{\theta}) \in \mathcal{F}_1\}$ is $\mathcal{C}_M^\alpha(\mathcal{X})$. Then for any $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$, we have

$$N(\varepsilon_1 C_L + \varepsilon_2, \mathcal{F}_1, \|\cdot\|_{\mathcal{F}}) \leq N(\varepsilon_1, \Theta, \|\cdot\|) \times \sup_{\theta\in\Theta} N(\varepsilon_2, \mathcal{C}_M^\alpha(\mathcal{X}), \|\cdot\|_\infty).$$

Proof. Let $\theta_1, \ldots, \theta_p$ denote an ε_1-net for $(\Theta, \|\cdot\|)$ with the additional restriction that $\theta_1, \ldots, \theta_p \in \Theta$, and for each θ_i, let $f_{i1}(z, \theta_i), \ldots, f_{iq_i}(z, \theta_i)$ denote an ε_2-net for the subclass $\{f(z, \theta_i) \in \mathcal{F}_1\}$ with a norm $\|\cdot\|_\infty$. Then note that for any $f(z, \theta) \in \mathcal{F}_1$, there exist θ_i and $f_{ij}(z, \theta_i)$ such that

$$\begin{aligned}\|f(z, \theta) - f_{ij}(z, \theta_i)\|_\infty &\leq \|f(z, \theta) - f(z, \theta_i)\|_\infty + \|f(z, \theta_i) - f_{ij}(z, \theta_i)\|_\infty \\ &\leq \varepsilon_1 C_L + \varepsilon_2.\end{aligned}$$

This proves the lemma since for each θ_i, the subclass $\{f(z, \theta_i) \in \mathcal{F}_1\}$ is $\mathcal{C}_M^\alpha(\mathcal{X})$. □

It follows from Assumption 3.4 that with probability approaching one,

$$\left\| \hat{f}_n(\cdot, \theta_1) - \hat{f}_n(\cdot, \theta_2) \right\|_\infty \leq C_L \left\| \theta_1 - \theta_2 \right\| \tag{B.4}$$

with some finite constant C_L. Hence, $\hat{f}_n(\cdot, \theta) \in \mathcal{F}_1$ with probability approaching one. Therefore, we can restrict the parameter space of $f(\cdot, \theta)$ to be $\mathcal{F}_1$.

To deal with $S_{n3}(\theta, f)$, consider a class of functions $\mathcal{M}_{\delta_1, \delta_2}$

$$\mathcal{M}_{\delta_1,\delta_2} = \{R(z, \theta, f) : (\theta, f) \in \Theta \times \mathcal{F}_1, \|\theta - \theta_0\| < \delta_1, \text{ and } \sup_{\{\theta : \|\theta - \theta_0\| < \delta_1\}} \|f(\cdot, \theta) - f_0(\cdot, \theta_0)\|_\infty < \delta_2\},$$

where $\mathcal{F}_1$ is defined in Lemma B.2. Then by Assumption 3.2 (a), an envelope function M_{δ_1,δ_2} for the class $\mathcal{M}_{\delta_1,\delta_2}$ has the form

$$M_{\delta_1,\delta_2} = (\delta_1 + \delta_2)\dot{m}(z, \delta_1, \delta_2).$$

Let $\|M_{\delta_1,\delta_2}\|_{L^2(P)} = [\int [M_{\delta_1,\delta_2}]^2 dP]^{1/2}$, where P is the probability measure of data Z.

Lemma B.3.

$$E\left[\sup_{\mathcal{M}_{\delta_1,\delta_2}} |S_{n3}(\theta, f)|\right] \leq Cn^{-1/2}(\delta_1 + \delta_2)\left(\delta_1^{\alpha_1} + \delta_2^{\alpha_2}\right)$$

Proof. By Lemma B.1, there is a positive constant C such that

$$E\left[\sup_{\mathcal{M}_{\delta_1,\delta_2}} \left| n^{1/2} S_{n3}(\theta, f) \right|\right] \leq C J_{[]}(1, \mathcal{M}_{\delta_1,\delta_2}, L^2(P)) \left\| M_{\delta_1,\delta_2} \right\|_{L^2(P)}. \tag{B.5}$$

First, note that by Assumption 3.2 (b),

$$\left\| M_{\delta_1,\delta_2} \right\|_{L^2(P)} \leq C(\delta_1 + \delta_2)\left(\delta_1^{\alpha_1} + \delta_2^{\alpha_2}\right).$$

Thus, to prove the lemma, it suffices to show $J_{[]}(1, \mathcal{M}_{\delta_1,\delta_2}, L^2(P)) < \infty$. Since $R(z, \theta, f)$ is Lipschitz in the parameters (θ, f) by Assumption 3.2 (a), we have, as in Theorem 2.7.11 of Van der Vaart and Wellner (1996, p.164),

$$N_{[]}(2\varepsilon \left\| \dot{m}(z, \delta_1, \delta_2) \right\|_{L^2(P)}, \mathcal{M}_{\delta_1,\delta_2}, L^2(P)) \leq N\left(\varepsilon, \Theta_{\delta_1} \times (\mathcal{F}_{\delta_1,\delta_2} \cap \mathcal{F}_1), \|\cdot\|_{\Theta \times \mathcal{F}}\right).$$

Then since $\left\| M_{\delta_1,\delta_2} \right\|_{L^2(P)} = (\delta_1 + \delta_2) \left\| \dot{m}(z, \delta_1, \delta_2) \right\|_{L^2(P)}$, substituting $\varepsilon(\delta_1 + \delta_2)/2$ for ε in

the both sides of the inequality above gives

$$
\begin{aligned}
&N_{[]}(\varepsilon \|M_{\delta_1,\delta_2}\|_{L^2(P)}, \mathcal{M}_{\delta_1,\delta_2}, L^2(P)) \\
&\quad \leq N\left(\varepsilon(\delta_1+\delta_2)/2, \Theta_{\delta_1} \times (\mathcal{F}_{\delta_1,\delta_2} \cap \mathcal{F}_1), \|\cdot\|_{\Theta\times\mathcal{F}}\right) \\
&\quad \leq N\left(\varepsilon(\delta_1+\delta_2)/4, \Theta_{\delta_1}, \|\cdot\|\right) \times N\left(\varepsilon(\delta_1+\delta_2)/4, (\mathcal{F}_{\delta_1,\delta_2} \cap \mathcal{F}_1), \|\cdot\|_{\mathcal{F}}\right) \\
&\quad \leq N\left(\varepsilon\delta_1/4, \Theta_{\delta_1}, \|\cdot\|\right) \times N\left(\varepsilon\delta_2/4, (\mathcal{F}_{\delta_1,\delta_2} \cap \mathcal{F}_1), \|\cdot\|_{\mathcal{F}}\right) \\
&\quad = N\left(\varepsilon/4, \delta_1^{-1}\Theta_{\delta_1}, \|\cdot\|\right) \times N\left(\varepsilon/4, \delta_2^{-1}(\mathcal{F}_{\delta_1,\delta_2} \cap \mathcal{F}_1), \|\cdot\|_{\mathcal{F}}\right) \\
&\quad \leq N\left(\varepsilon/4, \Theta, \|\cdot\|\right) \times N\left(\varepsilon/4, \mathcal{F}_1, \|\cdot\|_{\mathcal{F}}\right) \\
&\quad \leq N\left(\varepsilon/4, \Theta, \|\cdot\|\right) \times N\left(\varepsilon/(8C_L), \Theta, \|\cdot\|\right) \times \sup_{\theta\in\Theta} N\left(\varepsilon/8, \mathcal{C}_M^\alpha(\mathcal{X}), \|\cdot\|_\infty\right),
\end{aligned} \tag{B.6}
$$

where the last inequality follows from Lemma B.2. By Theorem 2.7.1 of Van der Vaart and Wellner (1996, p.155), there exists a constant C_K depending only on M, α, diam$\mathcal{X}$, and d_1 (recall that d_1 is the dimension of $\mathcal{X}$) such that

$$
\log N\left(\varepsilon, \mathcal{C}_M^\alpha(\mathcal{X}), \|\cdot\|_\infty\right) \leq C_K \left(\frac{1}{\varepsilon}\right)^{d_1/\alpha}. \tag{B.7}
$$

Then it is straightforward to verify that $J_{[]}(1, \mathcal{M}_{\delta_1,\delta_2}, L^2(P)) < \infty$ using the results obtained in (B.6) and (B.7). □

B.2 Asymptotic expansion of $S_{n2}(f(\cdot,\theta))$

Write $S_{n2}(f(\cdot,\theta)) = S_{n21}(f(\cdot,\theta)) + S_{n22}(f(\cdot,\theta_0))$, where

$$
\begin{aligned}
S_{n21}(f(\cdot,\theta)) &= n^{-1}\sum_{i=1}^n \Delta_{2i}[f(\cdot,\theta) - f(\cdot,\theta_0)] - \Delta_{20}^*[f(\cdot,\theta) - f(\cdot,\theta_0)] \\
S_{n22}(f(\cdot,\theta_0)) &= n^{-1}\sum_{i=1}^n \Delta_{2i}[f(\cdot,\theta_0) - f_0(\cdot,\theta_0)] - \Delta_{20}^*[f(\cdot,\theta_0) - f_0(\cdot,\theta_0)].
\end{aligned}
$$

Notice that the second term $S_{n22}(f(\cdot,\theta_0))$ does not depend on θ, therefore we can ignore this term. To establish an asymptotic expansion of the first term, further write

$$
\begin{aligned}
S_{n21}(f(\cdot,\theta)) = n^{-1}\sum_{i=1}^n &[\Delta_{2i} - \Delta_{20}^*][f_0(\cdot,\theta) - f_0(\cdot,\theta_0)] \\
&+ n^{-1}\sum_{i=1}^n [\Delta_{2i} - \Delta_{20}^*]\left[L_f(\cdot,\theta)\right],
\end{aligned}
$$

where $L_f(\cdot,\theta) = [f(\cdot,\theta) - f(\cdot,\theta_0)] - [f_0(\cdot,\theta) - f_0(\cdot,\theta_0)]$.

To deal with the second term of $S_{n21}(f(\cdot,\theta))$, consider a class of functions $\mathcal{L}_{\delta_1,\delta_2,\delta_3}$

$$\mathcal{L}_{\delta_1,\delta_2,\delta_3} = \Big\{\Delta_{20}[z, L_f(\cdot,\theta)] - \Delta^*_{20}[L_f(\cdot,\theta)] : (\theta, f) \in \Theta \times \mathcal{F}_{\delta_3}, \\ \|\theta - \theta_0\| < \delta_1, \text{ and } \sup_{\{\theta:\|\theta-\theta_0\|<\delta_1\}} \|f(\cdot,\theta) - f_0(\cdot,\theta_0)\|_\infty < \delta_2\Big\},$$

where $\mathcal{F}_{\delta_3}$ is a class of functions $f : \mathcal{Z} \times \Theta \mapsto \mathbf{R}^{d_f}$ that are in $\mathcal{F}_1$ and in addition, for any $\delta_3 > 0$,

$$\|[f(\cdot,\theta) - f(\cdot,\theta_0)] - [f_0(\cdot,\theta) - f_0(\cdot,\theta_0)]\|_\infty \le \delta_3 \|\theta - \theta_0\|. \tag{B.8}$$

Then by Assumption 3.4, for any $\delta_3 > 0$, $\hat{f}_n(\cdot,\theta) \in \mathcal{F}_{\delta_3}$ with probability approaching one. Therefore, we can restrict further the parameter space of $f(\cdot,\theta)$ to be $\mathcal{F}_{\delta_3}$.

Since $\sup_\theta \|L_f(\cdot,\theta)\|_\infty \le \delta_1\delta_3$, an envelope function $L_{\delta_1,\delta_2,\delta_3}$ for the class $\mathcal{L}_{\delta_1,\delta_2,\delta_3}$ is

$$L_{\delta_1,\delta_2,\delta_3} = C(z)\delta_1\delta_3 \tag{B.9}$$

for some $C(z)$ satisfying $\|\Delta_{20}[z, L_f(\cdot,\theta)]\| \le C(z)\|L_f(\cdot,\theta)\|_\infty$ for any θ.

Lemma B.4.

$$E\left[\sup_{\mathcal{L}_{\delta_1,\delta_2,\delta_3}} \left| n^{-1}\sum_{i=1}^n [\Delta_{2i} - \Delta^*_{20}]\,[L_f(\cdot,\theta)]\right|\right] = o\left(n^{-1/2}\delta_1\right) + o\left(n^{-1}\right).$$

Proof. As in (B.5), there is a positive constant C such that

$$E\left[\sup_{\mathcal{L}_{\delta_1,\delta_2,\delta_3}} \left| n^{-1/2}\sum_{i=1}^n [\Delta_{2i} - \Delta^*_{20}]\,[L_f(\cdot,\theta)]\right|\right] \le C J_{[]}(1, \mathcal{L}_{\delta_1,\delta_2,\delta_3}, L^2(P))\,\|L_{\delta_1,\delta_2,\delta_3}\|_{L^2(P)}. \tag{B.10}$$

Note that for any $\varepsilon > 0$,

$$N_{[]}\left(C[\|L_{\delta_1,\delta_2,\delta_3}\|_{L^2(P)}]\varepsilon, \mathcal{L}_{\delta_1,\delta_2,\delta_3}, L^2(P)\right) = N_{[]}\left(\varepsilon, [\|L_{\delta_1,\delta_2,\delta_3}\|_{L^2(P)}]^{-1}\mathcal{L}_{\delta_1,\delta_2,\delta_3}, L^2(P)\right) \\ \le N_{[]}\left(\varepsilon, \mathcal{L}, L^2(P)\right).$$

where $\mathcal{L}$ is a class of functions such that

$$\mathcal{L} = \Big\{\Delta_{20}[z, L_f(\cdot,\theta)] - \Delta^*_{20}[L_f(\cdot,\theta)] : (\theta, f) \in \Theta \times \mathcal{F}\Big\}.$$

By arguments similar to those used in the proof of Lemma B.2, for any $\varepsilon_1 > 0$ and $\varepsilon_2 > 0$,

$$N_{[]}\left(C(\varepsilon_1 + \varepsilon_2), \mathcal{L}, L^2(P)\right) \le N\left(\varepsilon_1, \Theta, \|\cdot\|\right) \times \sup_{\theta\in\Theta} N\left(\varepsilon_2, \mathcal{C}^\alpha_M(\mathcal{X}), \|\cdot\|_\infty\right)$$

for some C. It follows that $J_{[]}(1, \mathcal{L}_{\delta_1,\delta_2,\delta_3}, L^2(P)) < \infty$, provided that $\alpha > d_1/2$. Then the lemma follows immediately. □

Lemma B.5.

$$S_{n2}(\hat{f}_n(\cdot,\theta)) = n^{-1}\sum_{i=1}^{n}[\Delta_{2i} - \Delta_{20}^*][\partial f_0(\cdot,\theta_0)/\partial\theta^T](\theta-\theta_0) + o_p(n^{-1/2}\delta_1) + o_p(\delta_1^2) + R_{S_{n2}}$$

uniformly over $\theta \in \Theta_{\delta_1}$*, where* $R_{S_{n2}}$ *is a term that is independent of* θ*.*

Proof. The lemma follows immediately from Lemma B.4 since

$$\begin{aligned} &n^{-1}\sum_{i=1}^{n}[\Delta_{2i} - \Delta_{20}^*][f_0(\cdot,\theta) - f_0(\cdot,\theta_0)] \\ &= n^{-1}\sum_{i=1}^{n}[\Delta_{2i} - \Delta_{20}^*][\partial f_0(\cdot,\theta_0)/\partial\theta^T](\theta-\theta_0) + o_p\left(\|\theta-\theta_0\|^2\right). \end{aligned}$$

□

B.3 Asymptotic expansion of $S^*(\theta, f(\cdot,\theta))$

Define

$$\begin{aligned} H^*(\theta, f(\cdot,\theta)) = &\int_0^1 (1-s)\Big\{D_{ff}m^*(\theta, f_s(\cdot,\theta))[f(\cdot,\theta) - f_0(\cdot,\theta), f(\cdot,\theta) - f_0(\cdot,\theta)]\Big\}\,ds \\ &- \int_0^1 (1-s)\Big\{D_{ff}m^*(\theta_0, f_s(\cdot,\theta_0))[f(\cdot,\theta_0) - f_0(\cdot,\theta_0), f(\cdot,\theta_0) - f_0(\cdot,\theta_0)]\Big\}\,ds, \end{aligned}$$

where $f_s(\cdot,\theta) = f_0(\cdot,\theta) + s(f(\cdot,\theta) - f_0(\cdot,\theta))$.

Lemma B.6. *For any* $(\theta, f(\cdot,\theta))$ *in an open, convex neighborhood of* $(\theta_0, f_0(\cdot,\theta_0))$*,*

$$\begin{aligned} S^*(\theta, f(\cdot,\theta)) = &\frac{1}{2}(\theta-\theta_0)^T V_0(\theta-\theta_0) \\ &+ \frac{d}{d\theta^T}\Big(D_f m^*(\theta, f_0(\cdot,\theta))[f(\cdot,\theta) - f_0(\cdot,\theta)]\Big)\Big|_{\theta=\theta_0}(\theta-\theta_0) \\ &+ H^*(\theta, f(\cdot,\theta)) + o(\|\theta-\theta_0\|^2) + R_{S^*} \end{aligned}$$

uniformly over θ *in* Θ_{δ_1}*, where* R_{S^*} *is a term that is independent of* θ *and* V_0 *is defined in* (3.22)*.*

Proof. Write $S^*(\theta, f(\cdot,\theta)) = S_1^*(\theta) + S_2^*(\theta, f(\cdot,\theta))$, where $S_1^*(\theta) = m^*(\theta, f_0(\cdot,\theta)) - m^*(\theta_0, f_0(\cdot,\theta_0))$ and $S_2^*(\theta, f(\cdot,\theta)) = m^*(\theta, f(\cdot,\theta)) - m^*(\theta, f_0(\cdot,\theta))$.

First, consider $S_1^*(\theta)$. Since θ_0 is a unique minimizer of $m^*(\theta, f_0(\cdot,\theta))$ and θ_0 is in the interior of Θ (see Assumption 3.1 (a) and (b)), $dS_1^*(\theta)/d\theta = 0$. Then by simple calculus,

$$S_1^*(\theta) = \frac{1}{2}(\theta-\theta_0)^T V_0(\theta-\theta_0) + o(\|\theta-\theta_0\|^2), \tag{B.11}$$

where V_0 is defined in (3.22).

Now consider $S_2^*(\theta, f(\cdot,\theta))$. An application of Taylor's Theorem of $m^*(\theta, f(\cdot,\theta))$ around $(\theta, f_0(\cdot,\theta))$ (equivalently, evaluating (3.2) at $(\theta, f) = (\theta, f(\cdot,\theta))$ and $(\theta_0, f_0) = (\theta, f_0(\cdot,\theta))$ gives

$$\begin{aligned} S_2^*(\theta, f(\cdot,\theta)) &= D_f m^*(\theta, f_0(\cdot,\theta))[f(\cdot,\theta) - f_0(\cdot,\theta)] \\ &\quad + \int_0^1 \Big\{(1-s) D_{ff} m^*(\theta, f_s(\cdot,\theta))[f(\cdot,\theta) - f_0(\cdot,\theta), f(\cdot,\theta) - f_0(\cdot,\theta)]\Big\}\, ds, \end{aligned} \tag{B.12}$$

where $f_s(\cdot,\theta) = f_0(\cdot,\theta) + s(f(\cdot,\theta) - f_0(\cdot,\theta))$. By Assumption 3.6 (a), a Taylor expansion of the first term of the right hand side of (B.12) gives

$$\begin{aligned} D_f m^*(\theta, f_0(\cdot,\theta))[f(\cdot,\theta) - f_0(\cdot,\theta)] &= D_f m^*(\theta_0, f_0(\cdot,\theta_0))[f(\cdot,\theta_0) - f_0(\cdot,\theta_0)] \\ &\quad + \frac{d}{d\theta^T}\Big(D_f m^*(\theta, f_0(\cdot,\theta))[f(\cdot,\theta) - f_0(\cdot,\theta)]\Big)\Big|_{\theta=\theta_0}(\theta - \theta_0) \\ &\quad + R_{S_2}^*(\theta), \end{aligned}$$

where the Taylor series remainder term $R_{S_2}^*(\theta)$ is of order $o\left(\|\theta - \theta_0\|^2\right)$ because $f(\cdot,\theta)$ is restricted to be in a neighborhood of $f_0(\cdot,\theta)$. Thus, this result yields

$$\begin{aligned} S_2^*(\theta, f(\cdot,\theta)) &= \frac{d}{d\theta^T}\Big(D_f m^*(\theta, f_0(\cdot,\theta))[f(\cdot,\theta) - f_0(\cdot,\theta)]\Big)\Big|_{\theta=\theta_0}(\theta - \theta_0) \\ &\quad + H^*(\theta, f(\cdot,\theta)) + o\left(\|\theta - \theta_0\|^2\right) + R_{S^*} \end{aligned} \tag{B.13}$$

uniformly over θ in Θ_{δ_1}, where R_{S^*} is a term that is independent of θ, defined by

$$\begin{aligned} R_{S^*} &\equiv D_f m^*(\theta_0, f_0(\cdot,\theta_0))[f(\cdot,\theta_0) - f_0(\cdot,\theta_0)] \\ &\quad + \int_0^1 \Big\{(1-s) D_{ff} m^*(\theta_0, f_s(\cdot,\theta_0))[f(\cdot,\theta_0) - f_0(\cdot,\theta_0), f(\cdot,\theta_0) - f_0(\cdot,\theta_0)]\Big\}\, ds. \end{aligned}$$

The lemma now follows from (B.11) and (B.13). □

Combining the lemma above with Assumption 3.5 gives the following result.

Lemma B.7. *The following holds uniformly over θ in Θ_{δ_1}:*

$$\begin{aligned} S^*(\theta, \hat{f}(\cdot,\theta)) &= \frac{1}{2}(\theta - \theta_0)^T V_0 (\theta - \theta_0) \\ &\quad + \frac{d}{d\theta^T}\Big(D_f m^*(\theta, f_0(\cdot,\theta))[\hat{f}(\cdot,\theta) - f_0(\cdot,\theta)]\Big)\Big|_{\theta=\theta_0}(\theta - \theta_0) \\ &\quad + o_p\left(n^{-1/2}\|\theta - \theta_0\|\right) + o_p\left(n^{-1}\right) + o_p\left(\|\theta - \theta_0\|^2\right) + R_{S^*}, \end{aligned}$$

where R_{S^*} is a term that is independent of θ and V_0 is defined in (3.22).

B.4 Additional Proofs

Proof of Proposition 3.1. To verify Assumption 3.5, note that the left-hand side of the equation in Assumption 3.5 can be rewritten as $\hat{R}_{ff1}(\theta) + \hat{R}_{ff2}(\theta) + \hat{R}_{ff3}(\theta)$, where

$$\begin{aligned}
\hat{R}_{ff1}(\theta) &= \int_0^1 (1-s)\Big\{D_{ff}m^*(\theta, \hat{f}_s(\cdot,\theta)) - D_{ff}m^*(\theta_0, f_0(\cdot,\theta_0))\Big\} \\
&\quad \times [\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta), \hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)]\, ds, \\
\hat{R}_{ff2}(\theta) &= -\int_0^1 (1-s)\Big\{D_{ff}m^*(\theta_0, \hat{f}_s(\cdot,\theta_0)) - D_{ff}m^*(\theta_0, f_0(\cdot,\theta_0))\Big\} \\
&\quad \times [\hat{f}_n(\cdot,\theta_0) - f_0(\cdot,\theta_0), \hat{f}_n(\cdot,\theta_0) - f_0(\cdot,\theta_0)]\, ds, \\
\hat{R}_{ff3}(\theta) &= \int_0^1 (1-s)\Big\{D_{ff}m^*(\theta_0, f_0(\cdot,\theta_0))[\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta), \hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)] \\
&\quad - D_{ff}m^*(\theta_0, f_0(\cdot,\theta_0))[\hat{f}_n(\cdot,\theta_0) - f_0(\cdot,\theta_0), \hat{f}_n(\cdot,\theta_0) - f_0(\cdot,\theta_0)]\Big\}\, ds,
\end{aligned}$$

and $\hat{f}_s(\cdot,\theta) = f_0(\cdot,\theta) + s(\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta))$. Then it follows from (3.9) and one of conditions (b) (i)-(iii) that that $\hat{R}_{ffk}(\theta) = o_p\left(n^{-1}\right) + o_p\left(n^{-1/2}\left\|\theta - \theta_0\right\|\right)$ for $k = 1, 2$ uniformly over $\theta \in \Theta_{\delta_1}$.

Let $w_0(\cdot) = w(\theta_0, f_0(\cdot,\theta_0))$. By (3.9), write

$$\begin{aligned}
\hat{R}_{ff3}(\theta) &= \frac{1}{2}\int w_0(\cdot)\left\{[\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)]^2 - [\hat{f}_n(\cdot,\theta_0) - f_0(\cdot,\theta_0)]^2\right\} dP \\
&= \int w_0(\cdot)\left\{[\hat{f}_n(\cdot,\theta) - \hat{f}_n(\cdot,\theta_0)] - [f_0(\cdot,\theta)] - f_0(\cdot,\theta_0)]\right\} \\
&\quad \times \quad \left\{[\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)] + [\hat{f}_n(\cdot,\theta_0) - f_0(\cdot,\theta_0)]\right\} dP \\
&= \int w_0(\cdot)\left\{[\hat{f}_n(\cdot,\theta) - \hat{f}_n(\cdot,\theta_0)] - [f_0(\cdot,\theta)] - f_0(\cdot,\theta_0)]\right\}^2 \\
&\quad + \text{term not depending on } \theta.
\end{aligned}$$

Since condition (c) is satisfied,

$$|\hat{R}_{ff3}(\theta)| \leq o_p(\|\theta - \theta_0\|^2) + \text{term not depending on } \theta.$$

uniformly over $\theta \in \Theta_{\delta_1}$. Hence, we have proved the proposition. □

Proof of Proposition 3.2. It follows from conditions (a)-(d) and (3.12) that

$$
\begin{aligned}
&\frac{d}{d\theta^T}\left(D_f m^*(\theta, f_0(\cdot,\theta))[\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)]\right)\Big|_{\theta=\theta_0} \\
&= \frac{d}{d\theta^T}\left(\int [\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)] g(\cdot,\theta) dP\right)\Big|_{\theta=\theta_0} \\
&= \int \left[\frac{\partial \hat{f}_n(\cdot,\theta)}{\partial\theta} - \frac{\partial f_0(\cdot,\theta)}{\partial\theta}\right] g(\cdot,\theta) dP\Big|_{\theta=\theta_0} + \int [\hat{f}_n(\cdot,\theta) - f_0(\cdot,\theta)] \frac{\partial g(\cdot,\theta)}{\partial\theta} dP\Big|_{\theta=\theta_0} \\
&= n^{-1}\sum_{i=1}^n \int \tilde{\varphi}_{ni}(\cdot,\theta_0) g(\cdot,\theta_0)\, dP + \int \varphi_{ni}(\cdot,\theta_0) \frac{\partial g(\cdot,\theta_0)}{\partial\theta}\, dP + o_p(n^{-1/2}).
\end{aligned}
\tag{B.14}
$$

Then Assumption 3.6 is satisfied by condition (e). □

Proof of Theorem 3.4. Since this theorem can be proved by modifying the proof of Theorem 3.3, we will only indicate the differences that arise from the fact that $f(\cdot,\theta) \equiv f(\cdot)$. Abusing the notation a bit, we will use the same notation as in the proof of Theorem 3.3.

Re-define

$$R(z,\theta,f) = m(z,\theta,f(\cdot)) - m(z,\theta_0,f(\cdot)) - \Delta_1(z,\theta_0,f(\cdot))(\theta-\theta_0).$$

As shorthand notation, let $m_i(\theta,f) = m(Z_i,\theta,f(\cdot,\theta))$, $\Delta_{1i}(\theta,f) = \Delta_1(Z_i,\theta,f)$, and $R_i(\theta,f) = R(Z_i,\theta,f)$. Then $S_n(\theta,f)$ can be written as

$$S_n(\theta,f) = S_{n1}(\theta,f) + S_{n2}(f) + S_{n3}(\theta,f) + S^*(\theta,f),$$

where

$$
\begin{aligned}
S_{n1}(\theta,f) &= n^{-1}\sum_{i=1}^n \left[\Delta_{1i}(\theta_0,f) - \Delta_1^*(\theta_0,f)\right](\theta-\theta_0), \\
S_{n2}(f) &= n^{-1}\sum_{i=1}^n \left[m_i(\theta_0,f) - m_i(\theta_0,f_0))\right], \\
S_{n3}(\theta,f) &= n^{-1}\sum_{i=1}^n R_i(\theta,f) - R^*(\theta,f), \quad \text{and} \\
S^*(\theta,f) &= m^*(\theta,f) - m^*(\theta_0,f).
\end{aligned}
$$

Notice that by condition (c) of Assumption 3.7,

$$S_{n1}(\theta,\hat{f}_n(\cdot)) = n^{-1}\sum_{i=1}^n \left[\Delta_{1i}(\theta_0,f_0) - \Delta_1^*(\theta_0,f_0)\right](\theta-\theta_0) + o_p(n^{-1/2}\delta_1)$$

uniformly over $\theta \in \Theta_{\delta_1}$. Also, notice that $S_{n2}(f)$ can be ignored since this term does not depend on θ. The third term $S_{n3}(\theta, f)$ can be bounded in probability by $Cn^{-1/2}\delta_1^{1+\alpha_1}$ uniformly using arguments similar to those used to prove Lemma B.3. The last term $S^*(\theta, f)$ can be handled exactly the same as in Lemma B.7. Then the theorem can be proved by arguments identical to those used in the proof of Theorem 3.3 with $\varepsilon_n = n^{-1/2}\delta_1^{1+\alpha_1}$. □

Proof of Theorem 3.5. Since $\Gamma_1(z) \equiv 0$ in view of Assumption 3.6 and the assumption imposed here, this theorem is a direct consequence of Theorem 3.3. □

Lemma B.8. [Consistency for Example 2.1] *Assume that*

(a) θ_0 *is an interior point in* Θ*, which is a compact subset of* $\mathbf{R}^{d_\theta}$*,*

(b) $\Pr\left\{1(X \in \mathcal{T})[f_0(X_1 + X_2^T\theta, \theta) \neq f_0(X_1 + X_2^T\theta_0, \theta_0)]\right\} > 0$ *for every* $\theta \neq \theta_0$*, and*

(c) $\sup_{\theta\in\Theta} \left\| \hat{f}_n(\cdot, \theta) - f_0(\cdot, \theta) \right\|_\infty = o_p(1)$.

As $n \to \infty$, $\hat{\theta}_n \to_p \theta_0$.

Condition (b) is a high-level condition that imposes identification of θ_0 directly. Sufficient conditions can be found in Ichimura (1993, Assumption 4.2).

Proof of Lemma B.8. Define

$$\bar{S}_n(\theta) = n^{-1}\sum_{i=1}^n 1(X_i \in \mathcal{T})\rho_\tau\left[Y_i - \hat{f}_n(X_{1i} + X_{2i}^T\theta, \theta)\right] - n^{-1}\sum_{i=1}^n 1(X_i \in \mathcal{T})\rho_\tau\left(U_i\right), \tag{B.15}$$

where $U_i = Y_i - f_0(X_{1i} + X_{2i}^T\theta_0, \theta_0)$. To prove the theorem, it is more convenient to work with $\bar{S}_n(\theta)$ than (2.4). Write

$$\bar{S}_n(\theta) = \bar{S}_{n1}(\theta) + \bar{S}_{n2}(\theta),$$

where

$$\bar{S}_{n1}(\theta) = n^{-1}\sum_{i=1}^n 1(X_i \in \mathcal{T})\left\{\rho_\tau\left[Y_i - \hat{f}_n(X_{1i} + X_{2i}^T\theta, \theta)\right] - \rho_\tau\left[Y_i - f_0(X_{1i} + X_{2i}^T\theta, \theta)\right]\right\}$$

and

$$\bar{S}_{n2}(\theta) = n^{-1}\sum_{i=1}^n 1(X_i \in \mathcal{T})\rho_\tau\left[Y_i - f_0(X_{1i} + X_{2i}^T\theta, \theta)\right] - n^{-1}\sum_{i=1}^n 1(X_i \in \mathcal{T})\rho_\tau\left(U_i\right).$$

By the triangle inequality and condition (c),

$$\left|\bar{S}_{n1}(\theta)\right| \leq Cn^{-1}\sum_{i=1}^{n} 1(X_i \in \mathcal{T}) \left|\hat{f}_n(X_{1i} + X_{2i}^T\theta, \theta) - f_0(X_{1i} + X_{2i}^T\theta, \theta)\right| = o_p(1)$$

uniformly over $\theta \in \Theta$. By Lemma 2.4 of Newey and McFadden (1994, p.2129), $\bar{S}_{n2}(\theta)$ converges uniformly in probability to $S_0(\theta)$, where

$$S_0(\theta) = E\left[1(X \in \mathcal{T})\left\{\rho_\tau\left[Y - f_0(X_1 + X_2^T\theta, \theta)\right] - \rho_\tau(U)\right\}\right]. \tag{B.16}$$

It can be shown that $S_0(\theta)$ is uniquely minimized at $\theta = \theta_0$ using the identification condition directly imposed by condition (b). Therefore, the lemma can be proved by the standard consistency theorem for m-estimators (for example, Theorem 2.1 of Newey and McFadden (1994, p.2121)). □

References

Ahn, H. and C. F. Manski (1993): Distribution theory for the analysis of binary choice under uncertainty with nonparametric estimation of expectations, *Journal of Econometrics*, 56, 291-321.

Ahn, H. and J. L. Powell (1993): Semiparametric estimation of censored selection models with a nonparametric selection mechanism, *Journal of Econometrics*, 58, 3-29.

Ai, C. and X. Chen (2003): Efficient estimation of models with conditional moment restrictions containing unknown functions, *Econometrica*, 71, 1795-1843.

Aït-Sahalia, Y. (1994): The delta method for nonparametric kernel functionals, unpublished manuscript, University of Chicago.

Andrews, D. W. K. (1994a): Asymptotics for semiparametric econometric models via stochastic equicontinuity, *Econometrica*, 62, 43-72.

Andrews, D. W. K. (1994b): Empirical process methods in econometrics, in *the Handbook of Econometrics, Vol. 4*, ed. by R.F. Engle and D.L. McFadden. Amsterdam: North-Holland.

Andrews, D.W.K. (1995): Nonparametric kernel estimation for semiparametric models, *Econometric Theory* 11, 560-596.

Berger, M. S. (1977): *Nonlinearity and Functional Analysis: Lectures on Nonlinear Problems in Mathematical Analysis*, New York: Academic Press.

Chaudhuri, P. (1991): Nonparametric estimates of regression quantiles and their local Bahadur representation, *Annals of Statistics* 19, 760-777.

Chaudhuri, P., K. Doksum, & A. Samarov (1997): On average derivative quantile regression. *Annals of Statistics* 25, 715-744.

Chen, X. (2005): Large sample sieve sstimation of semi-nonparametric models, forthcoming in *Handbook of Econometrics, Vol. 6*, eds J. Heckman and E. Leamer. Amsterdam: North-Holland.

Chen, X., O. Linton, and I. Van Keilegom (2003): Estimation of semiparametric models when the criterion function is not smooth, *Econometrica*, 71, 1591-1608.

Chen, X. and X. Shen (1998): Sieve extremum estimates for weakly dependent data, *Econometrica*, 66, 289-314.

Das, M., W. K. Newey and F. Vella (2003): Nonparametric estimation of sample selection models, *Review of Economics Studies*, 70: 33-58.

Heckman, J. J., H. Ichimura and P. Todd (1998): Matching as an econometric evaluation estimator, *The Review of Economic Studies*, 65, 261-294.

Horowitz. J. L. (1998): Bootstrap methods for median regression models, *Econometrica*, 66, 1327-1351.

Ichimura, H. (1993): Semiparametric least squares (SLS) and weighted SLS estimation of single-index models, *Journal of Econometrics*, 58, 71-120.

Ichimura, H. (2006): Computation of asymptotic distribution for semiparametric GMM estimators, unpublished manuscript, University of Tokyo.

Khan, S. (2001): Two stage rank estimation of quantile index models, *Journal of Econometrics*, 100, 319-355.

Klein, R. W., and R. H. Spady (1993): An efficient semiparametric estimator for binary response models, *Econometrica*, 61, 387-421.

Newey, W. K. (1994): The asymptotic variance of semiparametric estimators, *Econometrica*, 62, 1349-1382.

Newey, W. K. and D. L. McFadden (1994): Large sample estimation and hypothesis testing, in *the Handbook of Econometrics, Vol. 4*, ed. by R.F. Engle and D.L. McFadden. Amsterdam: North-Holland.

Pakes, A. and S. Olley (1995): A limit theorem for a smooth class of semiparametric estimators, *Journal of Econometrics*, 65, 295-332.

Pollard, D. (1984): Convergence of stochastic processes, Springer-Verlag: New York.

Pollard, D. (1991): Asymptotics for least absolute deviation regression estimators, *Econometric Theory*, 7, 186-199.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): Semiparametric estimation of index coefficients, *Econometrica*, 57, 1403-30.

Powell, J. L. (1994): Estimation of semiparametric models, in *the Handbook of Econometrics, Vol. 4*, ed. by R.F. Engle and D.L. McFadden. Amsterdam: North-Holland.

Robinson, P. M. (1988): Root-N consistent semiparametric regression, *Econometrica*, 56, 931-954.

Sherman, R. P. (1994): U-processes in the analysis of a generalized semiparametric regression estimator, *Econometric Theory*, 10, 372-395.

Van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, New York: Springer-Verlag.

White, H. (1981): Consequences and detection of misspecified nonlinear regression models, *Journal of the American Statistical Association*, 76, 419-433.

Zeidler, E. (1986): *Nonlinear Functional Analysis and its Applications I: Fixed-Point Theorems*, New York: Springer-Verlag.